



University of East London

School of Architecture, Computing and Engineering

Title

A Novel CMAUT-UML Framework for the Optimisation of Clinical
Information System (CIS) and Prediction of CVD Percentage Risk:

By

Aloysius Adotey Edoh Jnr.

A Thesis submitted in partial fulfilment of the requirement of the University of East London
for the degree of Doctor of Philosophy.

September 2013

Abstract

This research critically analyses the different types of clinical data representation used in modelling Clinical Information Systems (CIS) and their limitations. It identifies space complexity, information overload, performance degradation, erroneous data retrieval and transmission as some of the main challenges caused by inappropriate data representation. Literature reviewed, indicated that object-oriented Health Level 7 (HL7), Entity Attribute Value (EAV), Advanced ERD with XML, and ERD –FOL (First Order Logic) are some of the contemporary methods used in modelling and optimising CIS. However, these approaches do not address the space complexity and information overload issues because of the multi-dimensional, complex large-scale nature of clinical datasets. Therefore, this research proposes a unique framework that uses object-oriented (UML) technique and combinatorial multiple attribute utility theory (CMAUT) as a new clinical data re-representation. In the CMAUT framework, the human organs, their multiple attributes and relationships are modelled using classes. The attributes of each organ class are written as logical expressions using CMAUT concepts, which are linked to each other with logical connectors *AND* for complementary organs such as cardiovascular and *OR* for substitutable organs like kidneys.

The logical expressions are converted into mathematical format, which serves as the utility objective function that is optimised using linear programming method subject to a set of constraint matrix. The constraint matrix is generated by transforming the multiple attributes in the CMAUT expressions into algebraic expressions by applying an algorithm that uses unit matrix and Raman transformation table. The output of the framework gives a set of attribute values, which optimal value maximises the overall utility of the objective function in the combinatorial organs. The algorithm maps the resultant attribute values to the appropriate attributes of the organs to determine the optimal amount of data required to be retrieved for primary health care investigation. The framework retrieves and transmits only needed data for investigation thus reducing the information overload and space complexity in the CIS.

The framework was implemented using the MATLAB software and validated with clinical data from the cardiovascular disease survey in England report. Functionality test conducted, revealed that for complementary organs the space complexity is $\theta(n + 1)$ using the framework and $\theta(2n)$ without the framework. Substitutable organs gave an exponential expansion of $\theta(2^n)$ in both cases. Simulation conducted showed that the mean size of the data retrieved for investigation using the framework is 463.5 bytes as compared to 1216.6 bytes without it. Statistical tests carried out using the output data from the framework gave a p-value of 0.000. Hence the hypothesis that the amount of data required for primary care health investigation can be reduced when the clinical data is re-represented with UML/CMAUT and optimised using LP based algorithm is statistically significant. For hypertension disease, by converting the optimal values from the framework into percentages give results similar to the percentage risk of the user been hypertensive. The output values were benchmarked against Framingham web based heart risk calculators and statistically analysed. Hence, the novelty of the framework is that it can be used for optimising CIS, as a multi-attribute decision tool and as an epidemiological prediction model for detecting high blood pressure diseases.

Acknowledgements

This Thesis could not have been written without the endless support and invaluable friendship of the people mentioned in this acknowledgement. I would like to express my gratitude to all those who gave me the possibility to do this PhD.

Foremost, I would like to express sincere gratitude to my supervisory team, who are Dr. Chris Imafidon and Dr. Aaron Kans, for providing me the continuous guidance and encouragement throughout PhD programme.

I would also like to thank my excellent advisors, which include Prof Allan J Brimicombe, Dr. Arunachalam, Ramas and Paul Bombo for their helpful and useful advice.

Special thanks to all my friends, who were always there for me whenever I needed them.

Last but not the least, I am greatly indebted to God and my family for their understanding, patience and support during the entire period of my research work.

Table of Contents

Title of the Thesis.....	(i)
Abstract.....	(ii)
Acknowledgement.....	(iii)
Table of Contents.....	(iv)
Nomenclature.....	(ix)
List of Acronyms.....	(x)
List of Figures.....	(xi)
List of Tables.....	(xiii)

Table of Contents

Chapter 1: Introduction.....	1
1.0 Chapter Introduction:	1
1.1 Research Background.....	1
1.1.1 Big Data and Information Overload in Data Intensive Industries.....	2
1.1.2 Challenges of Big Data and Information Overload in CIS	4
1.2 Types of Information Overload Solutions.....	5
1.3 Existing Research Gap in Clinical Information System.....	6
1.4 Scope and Objectives	7
1.5 Structure of Thesis.....	9
Chapter 2: Literature Review - Information Overload in CIS	11
2.0 Introduction	11
2.1 Taxonomy of Health Information Systems	11
2.2 Architecture of Clinical Information Systems and CDSS	13
2.2.1. Clinical Decision Support Systems (CDSS).....	13
2.2.2 Examples of Clinical Information Systems Architecture	15
2.3 Benefits and Challenges of Medical Information Systems and CIS	18
2.3.1. Challenges and Issues in Clinical Information Systems.....	19
2.4 Information and Data Representation in CIS	21
2.4.1. Clinical Coding and Data Representation:.....	22
2.4.2 Challenges of Data Coding and Representation in CIS	29
2.5 State of the Art Solutions used for Information Overload.....	30
2.5.1 Temporary and Contemporary Solutions for Information Overload	30
2.5.2 Supplementary Optimisation Solutions used in CIS.....	31
2.5.3 Limitations of Existing CIS optimisation techniques	34
2.6 Summary	34

Chapter 3: Methodology for UML -CMAUT Optimisation Framework.....	35
3.0 Introduction	35
3.1 Clinical Decision Support System and Decision Making Models	35
3.1.1 Conventional Clinical Decision Making models	36
3.1.2. Contemporary Clinical Decision Making models and their Limitations	37
3.2. Research Gaps in Optimisation of Clinical Information Systems (CIS)	41
3.2.1 Aim and Objectives:	42
3.2.2 Research questions:	42
3.3 Methodology used for the Research	43
3.3.1. Domain scenarios for modelling the CMAUT Prediction Framework.....	45
3.4 Data Collection and Analysis	46
3.4.1 Data for modelling and simulation of CMAUT CVD Framework (HSE, 2006).....	47
3.4.2 Design and Implementation of the CMAUT Framework;	48
3.5. Application of Statistical and Quantitative Methods.....	51
3.5.1. Statistical Methods.....	51
3.5.2 Quantitative Methods.....	52
3.6 Validation and Verification of the CMAUT-CVD CIS Framework	55
3.7 Summary	57
Chapter 4: Optimization and Clinical Data Re-representation in CIS	58
4.0. Introduction	58
4.1 Modern Clinical Data Re-representation Methods.....	58
4.1.1 First Order Logic/Entity Relationship Diagram (FOL/ERD)	58
4.1.2 Entity Attribute Value/Class Relation (EAV/CR) and ER Model:.....	61
4.2 Current Techniques for modelling Clinical Decision Support Systems	65
4.3 Multi-Attribute Utility Theory-MAUT	67
4.3.1 Implementation of Multi-attribute Utility Theory:	67
4.3.2 Challenges of Multiple Attribute Utility Theory;	69
4.4 Proposed new Clinical Data Re-representation using UML and CMAUT	70
4.4.1 The New Clinical Data Re-representation Mechanism	70
4.4.2 New Clinical Data Formalisation using CMAUT for CDSS.....	73
4.4.3 Conversion of Multiple Attributes into Utility Unit in CMAUT.....	77
4.5 UML Clinical Data Re-Representation with CMAUT formalisation	79
4.5.1 UML Clinical Data Re-representation of Kidney Diseases:.....	79
4.5.2 UML Clinical Data Re-representation of Cardiovascular Diseases (CVD):	81
4.6. Summary	83

Chapter 5: CMAUT Optimization Framework for CVD Risk Diagnosis:	84
5.0 Introduction	84
5.1 CIS Optimization Framework.	84
5.2 Data Re-Representation Mechanism using UML	86
5.3 Application Cardiovascular Disease (CVD) in the CMAUT Framework	88
5.3.1 The CMAUT Optimisation Algorithm:	90
5.3.2 Determination of Initial Clinical Absolute Percentage Risk (APR) in CVD	91
5.4 Modelling of the CMAUT Framework	92
5.5 Implementation of CMAUT Optimisation Diagnosis Framework models 1 and 2	95
5.5.1 Determination of APR Risk using CMAUT CVD Framework Model 1	95
5.5.2 Determination of APR Risk using CMAUT CVD Framework Model 2	102
5.6 Validation of CMAUT framework using Prevalence and Kappa statistic	108
5.6.1 Prevalence Computation	108
5.6.2 Computation of Kappa statistic	111
5.7 Simulation Results for CMAUT Diagnosis Framework model 1 and 2	113
5.7.1 Simulation Results, Tables and Figures for CMAUT Diagnosis model 1	113
5.7.2 Simulation Results, Tables and Figures for CMAUT Diagnosis model 2	118
5.8 Summary:	122
Chapter 6: Space Complexity and Clinical Data Reduction in CMAUT Framework	123
6.0 Introduction	123
6.1 CMAUT Framework	123
6.1.1 Space Complexity and the Application of Mathematical Operation	124
6.2 The Mathematical Operation Procedures	126
6.2.1 Conversion of the CMAUT logical expressions into Set of Inequalities:	126
6.2.2 Algorithm for Conversion of CMAUT expression to Set of Inequalities	126
6.3 Generation of Constraints for Complementary organs using CMAUT	128
6.3.1 Generation of Constraints for Complementary organs with CMAUT	128
6.3.2 Generation of Constraints for Complementary organs with Non- CMAUT	131
6.4 Generation of Constraints for substitutable organs with CMAUT	134
6.4.1 Substitutable organs using CMAUT Framework	135
6.4.2 Constraints Generation for substitutable organs with Non- CMAUT	138
6.4.3 Summary of the Space Complexity for CMAUT Framework	141
6.5 Analysis of clinical data sizes before and after optimisation	141
6.5.1 Statistical Analysis of data size before and after optimisation with CMAUT	142
6.5.2 Statistical Analysis of the Results using Pair T-test in SPSS	143
6.6 Summary	148

Chapter 7: CMAUT CVD Risk Prognosis Framework.....	149
7.0 Introduction:	149
7.1 CVD Predictive Percentage Risk and CMAUT Prognosis framework	149
7.2 The CMAUT Prognosis Framework	151
7.3 Principle of Modelling CMAUT Prognosis Framework	155
7.4 Implementation of CMAUT CVD Prognosis Framework – Model 1 and 2	158
7.4.1. Implementation of CMAUT CVD Prognosis Framework – Model 1	158
7.4.2 Determination of CVD PPR using CMAUT Framework Model1	159
7.4.3 Implementation of CMAUT CVD Prognosis Framework – Model 2	169
7.4.4 Determination of CVD PPR using CMAUT framework Model 2.....	169
7.5 Summary:	179
Chapter 8: Simulation of PPR with Web CHD Risk Calculators and Framingham Algorithms	180
8.0 Introduction	180
8.1 Methodology used for the selection of the CHD web risk calculators	180
8.1.1 Determination of the 10 years PPR with Internet model 1:- NHS BlackHeath.....	182
8.1.2 Simulation of the 10 years PPR with Internet model 2:- Patient UK	189
8.2 Framingham Algorithm for the determination of PPR values.....	197
8.2.1 Original Framingham algorithm from USA	197
8.2.2 Framingham algorithm from British Perspective:	200
8.2.3 The International version (Zgibor et al., 2006):.....	202
8.3 Simulation Results for the Framingham Equations versions I, II and III:.....	203
8.4 Summary	208
Chapter 9: Evaluation and Discussion	209
9.0 Introduction	209
9.1 Evaluation of Literature Review	209
9.2 Clinical data representation using UML and CMAUT - Success criteria 1	210
9.2.1 CMAUT Diagnosis Framework for CVD Risk Prediction -Success criteria 2.....	211
9.2.2 Comparison of APR for CMAUT Model 1 and 2:	212
9.3 Performance Evaluation of CMAUT Optimisation Framework	214
9.3.1 Space Complexity Analysis of CMAUT and Non-CMAUT CIS.....	214
9.3.2 Clinical Data Sizes and T-test analysis of CMAUT Optimisation Framework.....	217
9.4 Risk Prediction with CMAUT Prognosis Framework - Success criteria 2	221
9.4.1 Comparison of Predictive Percentage Risk for CMAUT Models 1 and 2:.....	221
9.4.2. Benchmarking CMAUT Prognosis framework with other CVD Prediction Tools	224
9.4.3 Computations of kappa value for Framingham equations I and II:	228
9.5 Prediction models Accuracy with Sensitivity/Selectivity and AUC/ROC	230
9.5.1 Discriminatory ability of three prediction models using sensitivity/selectivity	230
9.5.2 Prediction accuracy of the Prediction Models using AUC/ROC:.....	233

9.6 Performance accuracy of Prediction models using Likelihood Ratio.	237
9.6.1 Comparison and Interpretation of Likelihood Ratio Graphs	238
9.6.2 Comparison of the two Internet CVD calculators and Framingham Equations	239
9.7 Summary	241
Chapter 10: Conclusions and Recommendations for Further Research.....	242
10. 0 Introduction:	242
10.1 Conclusions	242
10.2 Contribution to Knowledge	246
10.3 Further Work and Recommendations.....	248
References.....	250
Appendix.....	264
Appendix 3.0 for Chapter 3	264
Appendix 5.0 for Chapter 5	264
Appendix 6.0 for Chapter 6:.....	269
Appendix 7.0 for Chapter 7:.....	273
Appendix 8.0 for Chapter 8.....	278
Appendix 9.0 for Chapter 9.....	287

Nomenclature

OR	Logical Expression
AND	Logical Expression
NOT	Logical Expression
XOR	Logical Expression for Exclusive OR
\neg	Symbol for the NOT Logical Expression
\wedge	Symbol for the AND Logical Expression
\vee	Symbol for the OR Logical Expression
\sum	Symbol for the summation of the objective function
\equiv	Equivalences
\Rightarrow	Implications
T	Temperature
HR	Heart rate
P	Blood pressure
F	Frequency
R	Resistance
V	Volume
A	Amplitude
(U)	Utility unit
$f(U_i)$	Utility function
β_0	Beta Coefficient for the first parameter
β_i	Beta Coefficients for the binary logistic regression
κ	Kappa value
μ	Location parameter
σ	Scale (dispersion) parameter
(u)	Initial Absolute Risk

List of Acronyms

Absolute Percentage Risk (APR)

Analytical Hierarchy Process (AHP)

Area Under the Curve (AUC)

ASSIGN methodology

Cardio Vascular Disease (CVD)

Clinical Decision Support System (CDSS)

Clinical Information System (CIS)

Combinatorial Components with Multiple Attribute (CCMA)

Combinatorial Components with Single Attribute (CCSA)

Combinatorial Multiple Attribute Utility Theory (CMAUT)

Confidence Interval (CI)

Coronary Heart Disease (CHD)

Entity Attribute Value /Class Relationship (EAV/CR)

Entity Relationship Diagram and First Order Logic (ERD/FOL)

Health Information Systems (HIS)

Health Survey for England Report (HSE, 2006)

Linear Programming (LP)

Medical Information System (MIS)

Multiple Attribute Utility Theory (MAUT)

National Institute for Clinical Excellence (NICE)

Predictive Percentage Risk (PPR)

CVD Risk Prediction Model (QRISK)

Receiver Operating Characteristics (ROC)

Sepsis-related Organ Failure Assessment (SOFA)

Unified Modelling Language (UML)

List of Figures

Figure 2.1: – Taxonomy of Health Information Systems

Figure 2.2: ‘Framework for Clinical Information System (Velde, 2000)

Figure 2.3: The NHS Spine – NHS Medical information system architecture

Figure 2.4: Microsoft three logical layer reference architecture (Hsieh et al., 2012).

Figure 2.5: EMIS Architecture and Database at GP surgery (Funk et al., 2009).

Figure 2.6: Examples of CVD READ codes in EMIS format

Figure 2.7: The RIM and the classes in the HL7 (adapted from Taylor, 2003)

Figure 3.1: Expert System Layout and its Components Parts (Laudon and Laudon, 2011)

Figure 3.2: NN structure with connections, weights and output neurons (Duetsh et al., 1994)

Figure 3.3: The cardiovascular system and the main organs, (Guyton and Hall, 2006)

Figure 4.1: Clinical Data Re-representation of ICD using FOL/ERD (De Keizer et al., 2000a)

Figure 4.2: Clinical Data Re-representation of ICD using FOL/ERD formalism

Figure 4.3: SQL for data retrieval from Conventional ERD table

Figure 4.4: SQL for data retrieval from EAV Table

Figure 4.5: Framework for optimising CIS

Figure 4.6: New clinical Data Re-representation (Imafidon et al., 2009)

Figure 4.7: Clinical Data Re-representation of Kidney using UML

Figure 4.8: Re-representation of heart diseases with UML

Figure 5.1: Flowchart and activity diagram of the CMAUT Framework

Figure 5.2: Re-representation of CVD risk factors for CMAUT framework

Figure 5.3: Output screen of CIS_CMAUT framework from desktop.

Figure 5.4: Procedure for determining Logistic Regression in SPSS

Figure 6.1: Graph of number of organs against number of constraints generated;

Figure: 6.2: Shows a comparison of Non – CMAUT using complementary organs (*AND*).

Figure 6.3: Non – CMAUT CIS statement using substitute OR organs

Figure 6.4: Percentage risk for heart disease calculation on the text file.

Figure 7.1: Procedure for determining Logistic Regression in SPSS.

Figure 7.2: Computation of PPR for CMAUT Prognosis model 1

Figure 7.3: MATLAB code for CMAUT Prognosis model 1

Figure 7.4: Output screen of CIS_CMAUT framework with predictive time for simulation

Figure 7.6: Predictive 10-years percentage risk for 3645 participants based on CMAUT Model – I

Figure 8.1: NHS BlackHeath Website for Heart Disease Risk calculator

Figure 8.2: The output for participant one from the simulation NHS BlackHeath Website:

Figure 8.3: Maximum Likelihood ratio for Internet model II

Figure 8.4: Predictive Percentage Risks for 10-years for Framingham I – II – III

Figure 8.5: ROC/AUC for Framingham I – II – III

Figure 8.6: Sensitivity and specificity for Framingham I – II – III

Figure 8.7: Maximum Likelihood ratio for Framingham I – II – III

Figure 9.1: Comparison of APR for 3645 participants based on CMAUT Diagnosis Framework Model I-II

Figure: 9.2: Shows a comparison of Non – CMAUT and CMAUT CIS using complementary organs (*AND*).

Figure 9.3: Shows comparison of Non- CMAUT and CMAUT CIS using substitutable organs (*OR*)

Figure 9.4: Before (red curve) and after optimisation (blue) of patient record data file

Figure 9.5: Comparison of PPR for 3645 participants on CMAUT Prognosis Framework Model I-II

Figure 9.6: Comparison of PPRs for 3645 participants using Internet Calculators Model 1-II

Figure 9.7: Comparison of PPR of 3645 participants using Framingham Models 1-II

Figure 9.8: Comparison of Sensitivity and Specificity of PPRs for the 3645 Patient PIND.

Figure 9.9: Comparison of ROC/AUC of PPR values for the 3645 Patient PIND.

Figure 9.10: Comparison of the Maximum Likelihood Ratio of PPR for the 3645 Patient PIND.

List of Tables

Table 2.1: Types of Clinical Coding and Data Representation Methods

Table 3.1: The measurable Clinical data used in the research

Table 3.2: Table for computation of Cohen Kappa

Table 4.1: Conventional ERD schema - relational database design

Table 4.2: Entity Attribute Value (EAV) Database Design

Table 4.3: EAV/CR Database tables an example of the EAV schema

Table 5.1- The three participants used for illustration and the simulation exercises

Table 5.2: Demography and Clinical data used in this Thesis:

Table 5.3: Shows attributes values for organs

Table 5.4: Raw data of the first 30 participants

Table 5.5: Absolute percentage risks and attributes variable values for the first 30 participants

Table 5.6: Classification Table for the 4316 participants in Model 1

Table 5.7: Calculation of TPR, FPR, LRP, and LRN, for the CMAUT Model I

Table 6.1: Representation of logical relations with linear inequalities Source (Raman et al., 1991):

Table 6.2: CMAUT complementary organs (AND)

Table 6.3: Non – CMAUT based CIS using complementary organs:

Table 6.4: CMAUT using substitutable organs (OR) also called partial substitutable organs

Table 6.5: Non-CMAUT using substitutable organs (OR) - partial substitutable organs

Table 6.6: The Output results of the Paired Samples Statistics:

Table 6.7A: Data size for 10 participants before and after optimisation with CMAUT

Table 7.1: The attributes values for organs

Table 7.2: The CVD data for the participants used for the illustration and simulations

Table 7.3: The Variables in the Equation

Table 7.4A: Predicative Percentage Risks for 10 years for the first 30 participants

Table 7.5A: Calculation of TPR, FPR, LRP, and LRN, for the MATLAB Model I PPR for 10 years for the first 30 participants (from Model I 3645 data sets)

Table 8.1A: Raw data of the first 10 participants

Table 8.2A: Predicative Percentage Risks for 10 years of the first 10 participants based on Internet

Table 8.3A: Calculation of TPR, FPR, LRP, LRN, for the Internet Model I of the first 10 participants

Table 8.4A: PPR values for 10 years for the first 10 participants based on Internet Model – II

Table 8.5A: Calculation of TPR, FPR, LRP, and LRN, for the Internet Model – II for the first 10 participants

Table 8.6A: predicative percentage risks for 10 years for the first 30 participants based on Framingham equation model I – II – III (I – USA, II – International, III – UK)

Table 8.7A: Calculation of TPR, FPR, LRP, and LRN, for the for the first 30 participants based on Framingham equation model I – II – III (I – USA, II – International, III – UK)

Table 9.1: No of CMAUT and Non-CMAUT constraints for substitutable complementary

Table 9.2: Data size for first 30 participants before and after optimisation with CMAUT

Table 9.3: The Output of the Samples Statistics Results for the Paired Samples T-Test:

Table 9.4B: Comparison of Absolute Percentage Risk values from CMAUT models 1 and 2 using the first 10 participants

Table 9.5: The actual_agree_YES and NO in tabular form

Table 9.6: The actual_agree_YES and NO in tabular form

Table 9.7: The for actual_agree_YES and NO in tabular form

Table 9.8A: Comparison of PPR values from CMAUT models, Internet calculators and Framingham equations:

Table 9.9A: Comparison of TPR and FPR for CMAUT models, Internet calculators and Framingham equations:

Table 9.10A: Comparison of LRP and LRN for CMAUT models, Internet calculators and Framingham equations;

Chapter 1: Introduction

1.0 Chapter Introduction:

This chapter starts with an overview of how the modern computer technology enables companies to capture, accumulate and store excessive amount of data in different formats. The concept of accumulation of excessive amount of data in computer systems for current or future use is known as “Big Data” (Chen et al., 2012). The chapter explains about how the huge data that has been accumulated and stored can be used to add value to the organisation’s business processes and assist in decision making. It then discusses the application of Big Data in three data intensive industries, namely the financial institutions, health industries and social networking sites. The chapter argues that there are challenges associated with Big Data in life and mission critical applications. It focuses on information overload in CIS, which is life critical application and outlines some methods used to address these problems. It establishes that the contemporary approaches used to address the issues of information overload do not work and therefore this research proposes a new hybrid technique that uses UML to capture clinical data and formalise it with CMAUT to reduce space complexity in information overload. The optimisation framework can also be used to determine the percentage risk of users been hypertensive. The chapter also outlines the aim, the objectives and contribution to knowledge as well as the structure of the Thesis.

1.1 Research Background

In present era, computer technology and the Internet are the key drivers of day to day life. These computer technologies are used for capturing and processing raw data, which is converted into meaningful format known as information for different activities and purposes. The Internet is used for the transmission of data and exchange of information for all types of applications. The advent of these new technologies also allows computer users to capture and accumulate huge amount of data in different formats. For example, the gigantic amount of structured, semi structured and unstructured data that are accumulated by organisations over many years is known as Big Data (Zikopoulos et al., 2012).

According to IBM Report (2013), 2.5 quintillion bytes of data are created every day, this data comes from social media sites, computer devices that are used to collect climate information,

business transaction records, digital images and videos and mobile phone GPS signals among others. The report also suggested that 90% of the Big Data in the world was generated in the last two years, which confirms the fact that people are becoming data dependant because of the advent of new computers and Internet technologies. The Big Data system is complex in nature and it is an untapped data source, which is valuable to many institutions although the continuous storage of Big Data is associated with many challenges (Agarwal et al., 2011).

The main benefit of this huge untapped data source is the ability to extract value from this Big Data using analytical tools and data mining techniques to identify trends and patterns for business applications (Herodotou et al., 2011). The Big Data systems are used for decision making, marketing, and creating Decision Support Systems (DSS). The advent of new technologies enables users to capture different applications, store and retrieve huge data for business intelligence, e-business and e-commerce purposes. These analytical processes empower users to add value to their Big Data set (Chen et al., 2012). Software houses such as Oracle and Microsoft have developed platforms with tools that allow the Big Data to be integrated with existing Databases and Data Warehousing. Examples of Big Data platforms with analytical tools are Hadoop, MyNoSQL and Starfish, which are used to improve efficiency and cost effective data retrieval from the Big Data. These Big Data analytical tools also eliminate delay and add value to the organisation's operations (Russom, 2011).

In spite of the success stories associated with the advent of the new computer technologies, Internet and Big Data, they have also created many challenges that must be addressed in order to facilitate their efficient and effective usage. Some of the challenges are security issues, information highway and information overload, which leads to the accumulation of excessive voluminous amount of data and information, errors in data transmission and problems with data storage (Floridi, 2012).

1.1.1 Big Data and Information Overload in Data Intensive Industries

The Big Data issues are mainly associated with industries that use data intensive applications such as financial institutions, the health industries, insurance, social networking websites and electronic media houses. This research analyses Big Data and information overload in the financial institutions, health industries and social networking sites. These organisations

generate huge amounts of data, which must be kept for many years for data analysis, decision making and business transaction purposes (Bawden et al., 1999).

Examples of Social networking sites that enable users to capture their day to day activities in multimedia format, store and exchange them for social purposes are the Facebook, Twitter WhatsApp and YouTube. Again, these social networking and media websites use applications that enable the creation of complex images and multimedia games that require large storage space, complex data retrieval mechanisms and data transmission techniques. These popular social networking sites create Big Data that are used for marketing, identification of users' behaviour pattern and data retrieval purposes. However, since these activities are not life or mission critical applications therefore the accumulated data can be archived or deleted to prevent information overload (Billinghurst and Starner, 1999).

In the financial institutions, such as the Banks, the data of each customer, which include their personal details, their daily business transactions and other activities are kept for a period of time and then archived for future purposes. Similarly, other financial organisations such as insurance companies, also capture, transmit and store all customers' transactions for a long period of time for decision making and Business Intelligence purposes. However, as the data size increases some financial institutions apply the delete policy to prevent information overload and improve the performance of their computer systems because they are not mission critical applications (Van Velsen et al., 2013).

In the health sector, each patient's personal detail and information are recorded and kept during each visit to the health centre. Each patient's record are captured and stored as clinical report and or history in different formats for medical decision making. With the invention of many new techniques in the medical sector, complex images and data are also captured and stored for diagnoses and prognoses purposes. Unfortunately, although the accumulated amount of data continues to increase, the delete policy discussed in section 1.2 cannot be applied in the medical sector because this is a life critical environment (Van Velsen et al., 2013). This research focuses on information overload in medical and clinical data intensive applications and investigates the challenges in the medical environment with a view to propose a solution.

1.1.2 Challenges of Big Data and Information Overload in CIS

According to Fernandes et al., (2012) the definition of Big Data in Health Information System (HIS) is the collection of large data sets in different formats, which have the three (3) Vs namely the volume, velocity, and variety. In this research, Big Data is defined as the collection and accumulation of very large and complex data sets with different data formats, which is difficult to process using the traditional database management system and contemporary data retrieval techniques.

In Fernandes et al. (2012) definition, the volume refers to the rapid rate at, which the amount of collected data is growing and the need to develop techniques and software for processing and managing the Big Data. The term “velocity” stands for the increasing frequency at which, data is captured and exchanged and therefore create information overload. In healthcare, variety means the capture of different forms of data, which include text; scanned documents, email, patient record and images that are accumulated and stored in the Big Data system.

The issues identified and discussed in the above definitions can be observed in all medical and CIS applications. For example in the medical field many new inventions are been introduced therefore medics write and store patients information in electronic format, which include variety of images and different data formats. Currently, patients’ details and day to day reports are recorded and stored in different electronic format, which increases the volume of patients’ records. The use of Internet and computer networks in health care environment to exchange and retrieve medical record for disease management, health care delivery and research purpose confirms the speed of change of data in medical field (Driscoll et al., 2013).

Again, in all healthcare information systems, like other data intensive industries, the creation and accumulation of Big Data leads to gigantic data storage problems, security issues, overloaded computer traffic, information highway, information overload, data transmission and retrieval problems (Billinghurst and Starner, 1999). The two main challenges associated with Big Data are Information Highway and Information Overload (IBM Report, 2013).

Information Highway is the process of transmitting and receiving excessive processed data or information through the Internet, which leads to information overflow. The presence of information overflow in the Internet based working environment is mainly associated with

search engines, social media website and exchange of email (Goulding, 2001). The challenges of information highway include the creation and exchange of irrelevant messages and information that cause poor computer performance and storage problems.

According to Fernandes et al., (2012), the main issue with Big Data is to attempt to make sense out of the information overload that has been created by the accumulation of excessive amount of data. Another issue with Big Data and information overload is to provide new insights into the growing volumes and sources of data in order to answer business, operational, and clinical questions in future. Problems created by information overload include performance degradation, security, retrieval, storage space and time complexity challenges (Ho and Tang, 2001). This research addresses the challenges caused by Information overload in CIS but will not examine the issues associated with Big Data and Information Highway.

1.2 Types of Information Overload Solutions

To address the issues associated with Information Overload, two main solutions namely the contemporary and supplementary solutions are recommended (Ho and Tang, 2001). The contemporary solution that is also known as temporary solution uses hardware, software techniques and data delete policy to address problems associated with Information Overload. The delete policy specifies that users must delete all files within a specific period else the computer system will automatically delete and or archive the files (Bertot, 2013). This method prevents the creation of information overload on the company's computing system. These solutions are temporary because as the volume of the accumulated data size increases the solutions are not able to resolve the Information Overload problems (IBM Report, 2013).

The second approach is known as supplementary solution. This involves the use of data aggregation and database optimisation techniques, the application of XML with user profile techniques, and the introduction of standard Clinical Client/Server Architecture (Velde, 2000). The data aggregation method uses advance SQL and data retrieval technique to store and search for information from big databases. These approaches are discussed in chapter 2 section 2.5.2 but they do not resolve issues associated with of information overload (IBM Report, 2013). The proposed Clinical Client/Server Architecture is made up of standard

medical application software, which is divided into three tiers namely; the view, domain and backend layers. This approach ensures that optimisation is carried out at the backend layer, which is the database. However, database optimisation and the application of object oriented (O-O) database do not address the information overload problem because the introduction of O-O models create an increase in the volume of data in the backend (Silberschatz, 2001).

The user's profile and XML technique captures and stores data in the user's requirement format that allows user specific information to be retrieved using XML techniques (Park et al., 2003). The user's profile technique unlike the Clinical Client/Server Architecture and database system facilitate the storage of the user's full data but only allows the partial retrieval of the stored data. This is further discussed in chapter 2.

1.3 Existing Research Gap in Clinical Information System

Different researches have been conducted on optimisation of Clinical Information Systems; however they all focus on optimising the client interfaces using XML technique (Park et al., 2003). The user profile system facilitates the retrieve of only relevant information instead of the entire data, therefore minimising the amount of data to be retrieved and reducing information overload. Database optimisation and data aggregation techniques are used to address information overload in clinical databases, however these methods are not very efficient (Safran and Chute, 1995) (Silberschatz, 2001).

The alternative approach is the application of data representation techniques where clinical data are captured using coding and medical languages. Some of the medical coding techniques such as ICD, SNOMED, READ and UMLS are used to capture and store clinical data therefore facilitate the easy retrieval of data using search techniques (Coiera, 2003). This medical coding technique resolves the problem of clinical data standardisation and exchange of clinical information between healthcare computer systems. However, these medical languages cannot be used in heterogeneous computing environment; therefore the Health Level Seven (HL7) data representation technique was introduced. The HL7 uses object oriented methodology to capture and represent clinical data. In addition it addresses the issue of standardisation and interoperability but does not resolve the issue of information overload in in CIS (Taylor, 2006). This is further discussed in chapters 2 and 3.

The current approach is the use of data re-representation technique where information is captured and re-represented in a format that can be used to model the information system, (Haimowits et al., 1988). This data re-representation technique was applied by De Keizer et al., (2000b), where they developed an ER diagram with a First Order Logics as a new form of data re-representation after studying the different types of medical coding and languages used in CIS. This new re-representation technique has not been implemented and it is discussed in Chapter 4. The second method is the use of Entity Attribute Value (EAV) and the enhanced Object Oriented Class Relationship (EAV/CR). This technique has been researched on and implemented in some USA health institutions (Nadkarni, 2002). Although this is an effective clinical data re-representation technique, it has a lot of performance limitations which are discussed in Chapter 4. The aforementioned methods have inbuilt CDSS systems but they do not have seamless information retrieval system and do not have optimisation mechanism that reduces space complexity and addresses the problems associated with information overload.

1.4 Scope and Objectives

The goal of this research is to address the issue of information overload by extending the concept of data re-representation technique in order to capture clinical data using UML class model. The data from the class model is written in mathematical format, using Combinatorial Multi Attribute Utility Theory (CMAUT) and logical connectors that is optimised with Linear Programming (LP) algorithm to reduce information overload in CIS. This framework is also used as an epidemiological tool to determine the percentage risk of a user been hypertensive or not. The aim and hypothesis of this research are presented in chapter 3 section 3.2.1.

- Objectives:

To achieve the goal of this research the following objectives are used:

1. Conduct detailed research to establish the problems associated with information overload in clinical information systems.
2. Discuss the methodologies used to address the information overload challenges and confirm the application of CMAUT optimisation Framework as a solution.
3. Conduct research on the application of class model and CMAUT as new form of clinical data re-representation technique for the development of CDSS in CIS.

4. Develop Diagnosis and Prognosis CardioVascular Disease (CVD) frameworks for optimising CIS.
5. Conduct mathematical operations and simulations using the frameworks to establish the data reduction in the CIS and information overload.
6. Conduct series of simulations using the two frameworks to determine Absolute Percentage Risk (APR) and Predictive Percentage Risk (PPR) in CIS
7. Using the clinical data from HSE 2006 survey, compare the PPR results of participants with hypertension from the CMAUT framework with results from the CVD calculators and Framingham algorithm.
8. Discuss and conclude the research.

This research has created a new optimisation framework that uses UML class model to capture multiple attribute clinical data in problem disease domain as a new form of re-representation technique. The **proposed** framework uses a radical different approach and has made contributions to knowledge, which are discussed in chapter 10, section 10.3.

1.5 Structure of Thesis

This Thesis started with an overview of the problems encountered in the use of Computer and Internet technologies and the creation of Big Data and information overload in Health Information Systems (HIS). Chapter 2 focuses on CIS and presents a detailed literature review on the challenges of excessive data in CIS. Chapter 2 discusses the modern techniques used in CIS to address the information overload and their limitations. The reviewed literature highlights that these techniques, cannot be applied in Big Data and CIS to resolve the information overload issue therefore demands new optimisation techniques.

The methodology Chapter 3 summarises the gaps in the literature reviewed in Chapter2 and states the hypothesis which forms the basis of this Thesis. This Chapter also discusses the methodology used to prove the hypothesis, which include quantitative and statistical methods. Demographic and clinical data from the Health Survey for England report (HSE, 2006) are used for the research. The six success criteria used to validate the output of CMAUT optimisation framework and verify the hypothesis are also presented.

In chapter 4, the two clinical data re-representation techniques namely; the ERD/FOL and the EAV/CR data re-representation techniques are analysed. This is followed by a review of three contemporary Clinical Decisions Support Systems (CDSS), which are Outranking, Analytical Hierarchy Process (AHP) and Multi-attribute utility theory (MAUT). The Chapter finishes with an explanation of how the proposed UML-CMAUT framework is used to capture clinical data that can be optimised to reduce information overload.

Chapter 5 discusses the design and implementation of the Cardiovascular Diagnosis optimisation framework using the data re-representation concepts described in Chapter 4. The framework is verified by using the Kappa and Prevalence calculations, to determine the framework that should be selected for discussion in this research.

The Chapter 6 discusses the storage space complexity and data size reduction in the CIS, when the CMAUT Optimisation Framework is used. This is done by conducting statistical analysis of the clinical data sizes before optimisation and after optimisation using the CMAUT framework.

In Chapter 7, the design and implementation of CMAUT Prognostic optimisation framework are discussed. The validation and verification of the efficiency of the CMAUT Prognostic framework are discussed in this chapter.

Chapter 8 discusses the assessment of the CMAUT optimisation framework with two existing Web-based CVD risk prediction calculators and three Framingham algorithms. The outputs values from the risk predictors are presented in tabular and graphical formats with explanations, before the validation and verification of the Framework are discussed.

The Chapter 9 starts with a summary of success criteria and benchmarks used to prove the hypothesis of this research. This is followed by discussion on the research findings and review of the Clinical Data Re-representation using the proposed UML-CMAUT framework. The results from the CVD Internet based calculators and the Framingham equations were benchmarked against the results from the CMAUT Prognosis framework. This was followed by the discussion on the calibrations and discrimination analysis of each of the frameworks.

The results of the predictive percentage risk (PPR) values from the CMAUT Prognosis framework were compared with the PPR values from Framingham equations and CVD Risk calculators and analysed. The accuracy of the prediction models using sensitivity and specificity are discussed using the NICE criterion of 20% as the basis for the comparison. This is followed by the discussion and evaluation of the Area Under the Curve (AUC) and Likelihood Ratios of each of the CVD prediction models.

Finally, Chapter 10 outlines the summary of the research including the proposed recommendation and further works. Significant contributions to knowledge have also been summarised in this chapter.

Chapter 2: Literature Review - Information Overload in CIS

2.0 Introduction

This chapter presents an overview of the current state of the art of Clinical Information System (CIS). It discusses the different types of CIS architecture and the components that make up a CIS. It argues that Client Server System forms the basis of CIS and that the middle tier is also known as the application server. The middle tier serves as the Clinical Decision Support System (CDSS). From the discussion in Chapter 1, CIS captures different types of clinical data, stores and processes them for analysis and decision making. Therefore, the various forms of capturing clinical data are discussed in this chapter in order to identify their strengths and weaknesses. The Chapter also examines issues encountered in the application of the existing clinical data representation methods and how they are addressed. CIS challenges such as, information overload, performance degradation and erroneous data transmission in distributed CISs are discussed. The chapter finishes with an evaluation of the contemporary and supplementary methods used to address information overload issues.

2.1 Taxonomy of Health Information Systems

Health Information Systems (HIS) is defined as the application of computer technology to capture, store and retrieve medical data as well as automate health care business processes to deliver cost effective health care (Raymond and Dold, 2001). In order to deliver efficient health care, manage patients' records and satisfy the requirement of healthcare customers, all hospital business processes must be computerised. The hospital business processes that must be computerised include duty rotas, pathology, laboratory reports, epidemiology, e-perception, e-health (Page, 2012) (Steve, 2006). This research focuses on CIS, which is a subset of medical information systems as shown in Figure 2.1 below. CIS deals with the management and delivery of patient care as well as patients' medical needs. It focuses on the anatomy of the patient rather than the hospital settings and the health care infrastructure.

The Taxonomy of Health Information System (HIS) captures all the business activities, which are automated in the hospital environment (O'Carroll, 2003). In addition, HIS handles health care activities, such as laboratory tests, surgical and pharmaceutical activities. The block diagram for the taxonomy of Health Information Systems is presented below.

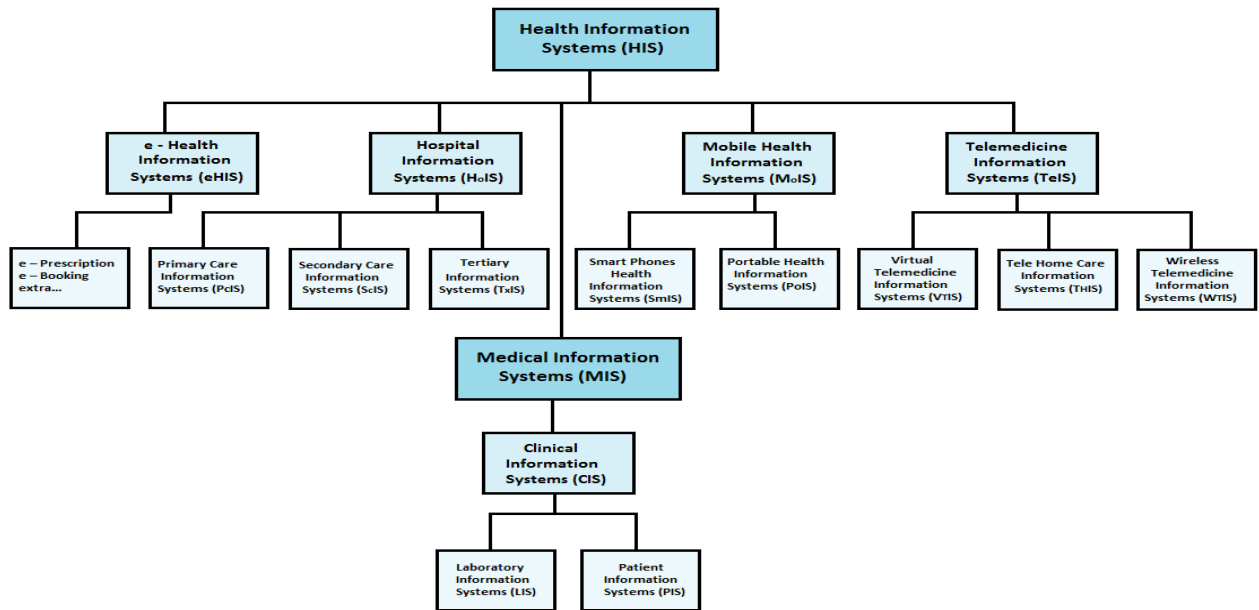


Figure 2.1: – Taxonomy of Health Information Systems (HIS).

In Figure 2.1, the Health Information System comprises of hospital information a system, which is made up of the primary care, secondary care and tertiary hospitals (aka university hospitals where medical practitioners are trained and medical research carried out. In UK, Primary care is also known as GP and it is the first point of call for all patients. Another subsystem in Health Information Systems is e-health, which deals with the application of Internet or web based health care delivery (Granger, 2006).

Electronic Health (e-Health) deals with health activities and procedures, which are presented over the Internet or use e-commerce infrastructure (Hsieh et al., 2012). E-Health systems use the Internet and web applications to implement health care, conduct medical investigations as well as manage and deliver efficient health care. Examples are e-referral, e-prescriptions and e-epidemiology. M-health, on the other hand is the application of mobile technology in the delivery of healthcare. These types of HIS are real time processes that require large amount of streamed data between medical centres (Hartvigsen and Pedersen, 2012). These examples show that there are many challenges in the use of Big Data and information overload in HIS application software as stated in section 1.1 (Fernandes et al., 2012).

2.2 Architecture of Clinical Information Systems and CDSS

In healthcare delivery there are different forms of computer system configurations that are tailored to HIS requirements as shown in Figure 2.1. There is the standalone CIS where clinical data are captured by interviewing the patient and the data recorded in text format using natural language, which is then stored as documents. Another type of computer configuration is the distributed CIS, which is used in the secondary and tertiary health information system (Mattmann, 2003). The distributed CIS allows GPs to link to the other centralised hospitals for information exchange. The third type of CIS, is the heterogeneous and integrated CIS, where hospitals and GPs with different hardware and software settings can link up with each other to share and exchange patient details. This system is used by the NHS in their CIS, and it is discussed in section 2.3 below (Velde, 2000).

In all these CIS configurations, the advent of new technology allows medics to capture, store and exchange accurate information on the patients' medical condition for efficient healthcare management. However, in some cases, the clinical data is captured using clinical images, such as digital image communication system also known as DiCom and PACs (Picture Archiving and Communications systems), which takes a huge amount of storage space and create huge processing overhead (Hoelzer et al., 2003). Many different methods have been used for the development of CISs with a view to address the above issues, but they are all not optimised, hence the issue of information overloads still exist.

2.2.1. Clinical Decision Support Systems (CDSS)

The architecture of CIS is made up of Client Server Systems (CSS), which consist of three layers (Velde, 2000). The medic interacts with the automated information system via an interface which is known as the application layer. The middle layer which is the logical layer or middle tier houses the CDSS. In the CDSS are the business processes that the CIS is required to execute in order to perform the different clinical functions that the system is designed for. In software terms the middle layer is also known as application server or middleware while the third layer is the storage area, where the patient's details and records are stored.

The CDSS layer stores the business logics or algorithms needed to execute the functionalities that the system is designed to perform (Massimo, 1998). These functionalities include clinical data retrieval, capture of the clinical data, data exchange, mapping of clinical data to clinical codes, booking clinical appointments and e-prescription are some of the functions performed by the middle layer. There are different types of CDSS, which include the use of Artificial Intelligence (AI) and automatic searchable indexes technique (Brownin et al., 2002). These CDSS techniques and their implementation are discussed in Chapters 3 and 4.

The CDSS is made up of two components, the logical component, which is the software and the physical component known as the hardware. To address the issue of information overload in CIS that is caused by the capture and storage of huge amount of clinical data, the contemporary hardware solution techniques are used. An alternative approach is to use distributed application servers, which allows different functions to be installed on different hardware at different sites. This is discussed in example 2.2.2 under the Standard Clinical Client/Server Architecture that has been proposed by Velde, (2000).

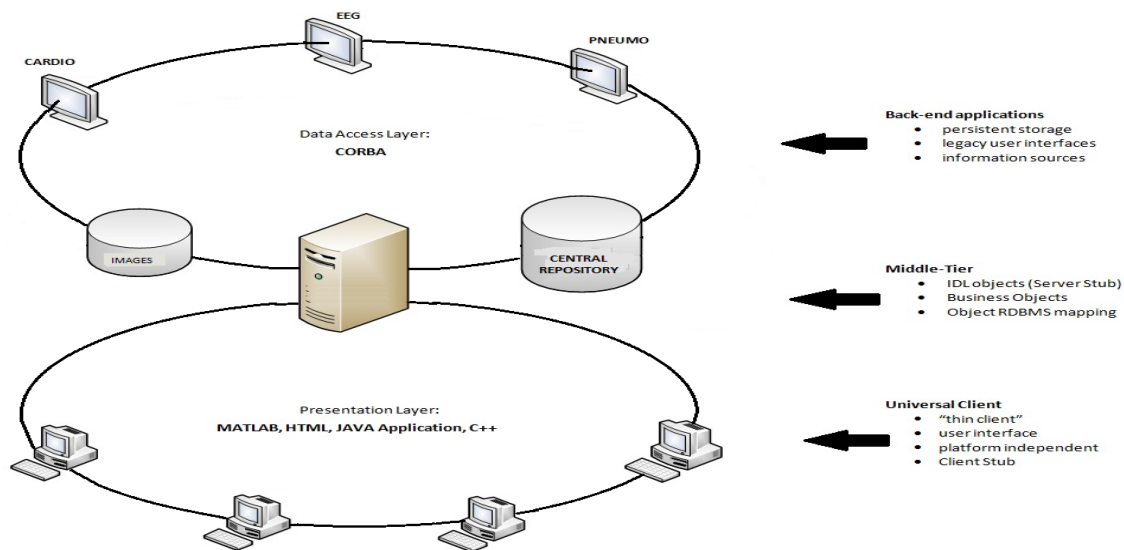


Figure 2.2: Framework for Clinical Information System (Velde, 2000).

The CIS architecture shown in figure 2.2 was proposed, to standardise the different types of CIS architecture, used in Europe and across the world. This is because the non-standard CIS architecture creates many problems such as, interoperability issues, erroneous data transfer, extendibility and scalability.

The proposed CIS standard architecture is based on Object Oriented Database that uses CORBA, as a middleware and HTML with Java application, as an interface. Images and other medical data are stored in the Object Oriented Database. This standard integrated architecture was implemented and the results were found to be efficient but have performance limitations, due to information overload from the accumulation of clinical data (Velde, 2000).

2.2.2 Examples of Clinical Information Systems Architecture

Below is a summary of the three examples of the types of architecture used for the development of CIS. These examples illustrate the nature of the non-standard CIS architecture used in healthcare delivery in UK with their limitations.

- Example 2.1: NHS Medical information system architecture:

In UK, the NHS has its own HIS architecture, which uses the three tier model. An example is the Figure 2.3 below. This HIS architecture used by NHS has an interface as the front end and the middle tier is Vision Software that integrates all heterogeneous GP CIS together. The backend is a database where all the medical information is stored. This system was designed to capture and integrate all hospitals and GPs in the UK (Spronk, 2007).

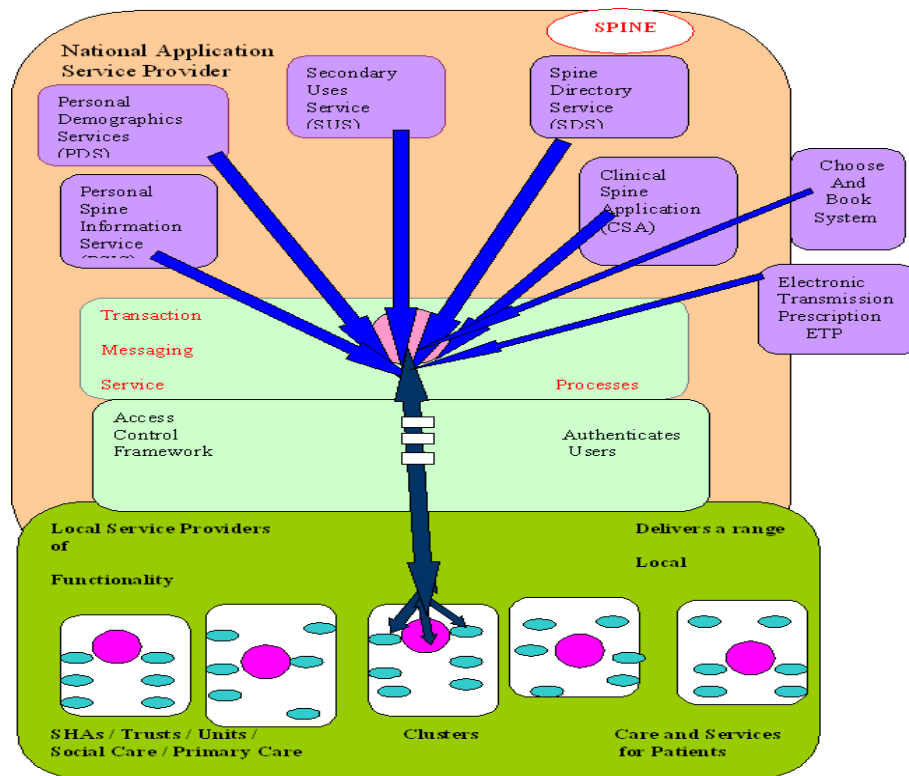


Figure 2.3: The NHS Spine – NHS Medical Information System Architecture (Spronk, (2007).

However, research conducted using the NHS architecture revealed that although it allows the CIS in the individual GPs to be connected to the secondary healthcare computer systems, it has performance limitations. These issues include network delay, error in data transmission and interoperability problems (Steve, 2006).

- Example 2.2: Microsoft Three logical layer reference architecture

The Microsoft Company has proposed three tier architecture for healthcare delivery (Hsieh et al., 2012). The Microsoft architecture in Figure 2.4 comprises of a user interface, which is a Microsoft browser, a middle tier that uses a dot NET middleware with SharePoint services and at the backend is the Microsoft database storage server. In this three logical layer reference architecture, the middle layer performs two functions: the provision of web services and integration of Microsoft system with other clinical information applications. This architecture is known as homogenous system, because all the components in it are Microsoft products therefore it cannot interoperate with other non- Microsoft Systems unless a third party software is added to the system (Hsieh et al., 2012).

This means any non-Microsoft application that needs to be integrated with the Microsoft architecture requires third party software to facilitate the integration. Therefore this architecture has interoperability and integration problems.

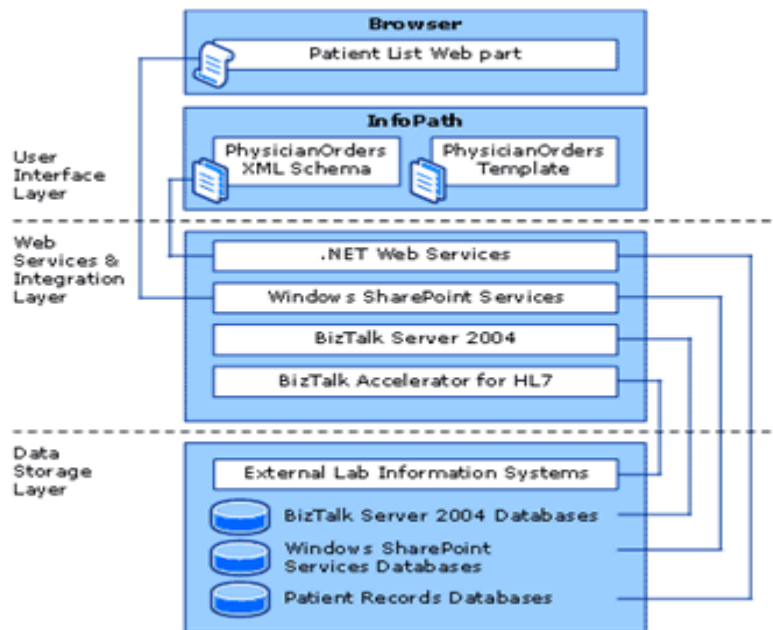


Figure 2.4: Microsoft three logical layer reference architecture (Hsieh et al., 2012).

- Example 2.3: EMIS Architecture for GP surgery

The third type of HIS architecture is the EMIS Solution (Electronic Medical Information System). The EMIS architecture Figure 2.5 is used by GPs in the UK. It is a standalone three tier architecture, which when installed in a clinical setup enables the users to have a user friendly interface. The middle tier has the software that performs all the basic functionalities that is required for the day to day operations in a GPs primary care set up. The backend is a Relational Database where all the patient records are kept. Although this EMIS is a standalone system, it can be integrated to the Spine of the NHS architecture described in Example 2.1, through the Internet connection known as Network 3 (N3) (Funk et al., 2009).

According to EMIS, their architecture can be converted into distributed system with optimised concurrency control and advanced fault tolerance systems and therefore makes the EMIS more reliable and easily scalable. The use of HL7 facilitates easy exchange of clinical information between the CIS and auxiliary systems (Funk et al., 2009).

However, experience users indicate that EMIS architecture has performance issues. Therefore NHS has recommended that the operation and installation of the EMIS Solution should be discontinued (Steve, 2006).

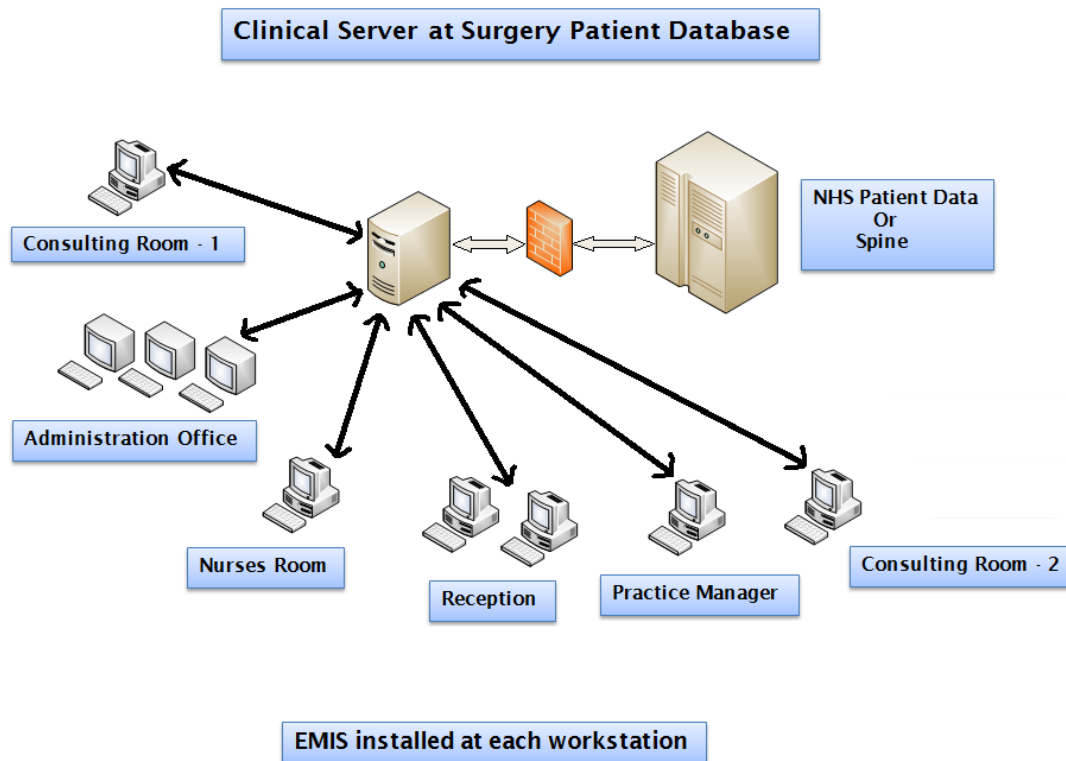


Figure 2.5: EMIS Architecture and Database at GP surgery (Funk et al., 2009).

2.3 Benefits and Challenges of Medical Information Systems and CIS

The Medical Information System covers all aspects of health care namely the public health, personal health, health care, disease prevention, eradication and epidemiology. The key benefits of medical informatics (Granger, 2006) include the following:

- The use of electronic medical records for patients instead of the paper based records.
- The secure electronic networks for the delivery of real-time medical data at anytime and anywhere the patient and or clinician need them.
- Electronic transmission of error free medical and clinical test results.
- Electronic prescribing of medications, treatments, and tests without errors.
- Decision Support Systems (DSS) that provide clinicians with latest information on the best practices and treatment options.

- The use of electronic devices such as handheld computers to make information available at the point of care to reduce space complexity.

In addition to the above, Medical Information System (MIS) aka medical informatics is crucial for improving the quality, safety, and effectiveness of the healthcare. This real-time and comprehensive clinical data provide accurate information for decision making rather than the existing paper-based systems (Steve, 2006). The use of electronic information system results in better medical treatment and decisions making with fewer medical errors. It encourages active involvement of patients themselves, resulting in more patient-specific and patient-centred care. This method leads to improvement in the quality of patient care and increases the productivity of clinicians. However, in spite of these benefits, CISs have challenges, which are discussed below (Granger, 2006).

2.3.1. Challenges and Issues in Clinical Information Systems

Clinical Information Systems suffer from a number of problems and challenges, which must be addressed by the systems of tomorrow (Steve, 2006). Some of the issues to be addressed include; accurate data exchange and seamless interoperability among all parties and computer systems in patient care (Bath, 2008). There is also a need to have interoperability between different HISs as well as improve the areas listed below.

- Develop Intelligent DSS and maintain all obsolete medical systems,
- Increase the domain data size and reduce rate of information exchange,
- Support domain experts to have direct control over the information system design and handle change management in health care delivery.

Although all countries are trying to reduce the healthcare costs, by restructuring the delivery of healthcare services, it is evident that, these challenges can only be achieved through effective information system. According to Ganger, (2006) the provision of right clinical information at the right time and place, to the right person will help to resolve the healthcare problems.

This research analyses CIS challenges which include: interoperability and data exchange security; data standardisation and comparability; data quality and erroneous data transmission and information overload and storage of huge amounts of clinical data, as discussed in section 2.6. Security issues are not addressed because it is beyond the scope of this research work (Anderson and Street, 1996).

- Interoperability and data exchange

Interoperability is the ability of one system to exchange data with another system regardless of their configuration, hardware and software constitution. Healthcare computer systems utilise many different information systems, both within an organization and across organizations. Physicians working in hospitals, also have different systems in their offices, but need to access data of their patients from the hospital. These different systems must be able to exchange data between themselves (Winsten, 1996).

To achieve interoperability between different information systems, standard messaging systems must be implemented (Taylor, 2006). The implementation of these standard systems is costly and time-consuming and therefore restricts their adoption by small GPs. To address the issue of interoperability in healthcare, semantic Web Services and ontology are proposed but Microsoft uses Health Level Seven (HL7) clinical standard (Werner et al., 2003).

- Data standardisation and Compatibility

In clinical application, data compatibility requires that the meaning of data is consistent when shared among different parties. Lack of compatible data can cause an impact on patient care (Taylor, 2006). Standard healthcare vocabularies ensure that data shared across systems are comparable at the most detailed level. However, MIS vendors and healthcare providers can create their detailed vocabularies, but it must conform to one or more of the standard clinical coding and languages discussed in section 2.4 below. For example, the EMIS group whose architecture was discussed in section 2.4 has proprietary set of terms, which is compatible with the READ code for all clinical areas and it is explained in section 2.4 (Coiera, 2003).

- Data Quality and Erroneous Medical Data

It is difficult to measure the quality of healthcare data, however, every provider points to cases where, data quality was suspected to be incorrect or could not be validated. Many CIS do not incorporate sufficient data editing capability, uniformity of the units of measurement and other controls into their CIS architecture. This leads to error in the data and loss of data for patient care (Pietra et al., 2005). Again, the absence of unique identifier on some patient's record leads to erroneous medical data storage, retrieval and data exchange. The Administrative Simplification Provisions of Health Insurance Portability and Accountability Act (HIPAA) recommend the use of unique identifier all patients' records.

However, there is concern that the introduction of unique identifier for patients will give easy access to the entire records of a particular patient (Coiera, 2003). The standards of medical vocabulary and coding are discussed in section 2.7 but standard message format and security systems are not covered in this research.

2.4 Information and Data Representation in CIS

According to Cimino et al. (2002), CIS is a branch of medical informatics that deals with the use of information technology for monitoring, diagnosing and analysing clinical problems in order to make decision and eradicate diseases. Advances in computer technology, has been deployed in many fields but limited applications can be found in clinical informatics. This is because of the complex structure and the nature of clinical data, which makes them sensitive, dynamic and multidimensional. For example, human organs have multiple attributes that can be measured in continuous variables and stored as real-time data for future use. Therefore the size of clinical data is gigantic and has space complexity issues (Rassinoux, 1998).

Two common methodologies used to capture clinical data, are the use of Natural Languages (NL) and Digital Image Communication System (DiCom) or Picture Archiving and Communications System (PACs) (Hoelzer, 2003). Each of these methods has its limitations. For example, the Natural Language (NL) uses two approaches, which are either the patient's data is captured by using manual handwriting that can be difficult to read or the use of computerised data processing method.

In both cases the data must be accurate and adhere to the standard Clinical Nomenclatures or coding and must be stored for future purposes. The DiCom and PACs allow medics to capture medical data in the form of images or pictures, although these methods give precise information, they create high processing overheads, storage space issue and complex retrieval problems (Cimino et al., 2002). Thus, these techniques used to capture clinical data, contribute to Big Data and information overload.

2.4.1. Clinical Coding and Data Representation:

To resolve the issues raised in section 2.3, standard medical terminologies, vocabularies and nomenclatures are developed and maintained by standardisation organizations. These clinical coding, Medical Language and clinical Data Representation include ICD-10, SNOMED, Unified Medical Language System (UMLS) and READ among others (Rodrigues, 2009).

These Clinical Coding/Language and Clinical Data Representation methods are used to capture and represent clinical data using High Level Languages or Natural Languages. The clinical coding ensures that there is standardisation in the use of medical terminologies and vocabularies, when clinical data are captured and exchanged between clinicians (Taylor, 2006). Most Data Representation systems have inbuilt clinical coding mechanism, that facilitate the mapping of medical concepts to the appropriate standard medical code, to ensure data compatibility between users and reduce errors in medical records (O'Carroll, 2003).

The clinical coding and Medical Languages use different data structures, to capture as well as represent and retrieve clinical data. Some of the data structures used for clinical coding/ languages are shown in Table 2.1 below. The advantages of using these data structures are; they are scalable, user friendly and provide easy to search facilities. However, as the size of the clinical coding or medical languages increase, the search and retrieval mechanisms become complex and less user friendly. Again, the advance coding language such as UMLS use **is_a** link to connect different medical and clinical concepts and therefore complex retrieval mechanisms are required to search for information in these systems (Taylor, 2006). These limitations and other challenges such as the introduction of Health Level Seven (HL7) to address the problems of interoperability are discussed under each of the five commonly used medical coding below.

1. ICD-10 from World Health Organization (WHO)

One of the significant international coding and classification system, known as International Statistical Classification of Diseases (ICD) was developed and maintained by the World Health Organisation (WHO). The ICD-10 stands for the International Statistical Classification of Diseases and Related Health Problems, 10th revision. This Tenth Revision is a series that was formalized in 1893, to record and analyse Diseases as well as list the International causes of Death (Manchikanti et al., 2013).

The data structure of the ICD Revision 10 comprises of a multiple-axis classification system with 21 chapters, which covers various diseases. The first volume contains the disease classification, which are located at the three-character and four-character levels. Other classifications are listed under special tabulation such as mortality, morbidity, definitions of medical concepts, and the nomenclature regulations. There is a special volume that includes report on the International Conference for the Tenth Revision, which covers the entire complex deliberations but could not be included in the ICD -10 coding. The Final volume has alphabetical index with introductory information on practical advice on how to make the best use of the index. To facilitate efficient coding and code searching, the index includes commonly used diagnostic terms as synonyms for the terms formally accepted for use in the classification (Asghari and Mahdavian, 2013).

The ICD is an international Clinical Coding system but different countries have modified the coding to meet their local medical and clinical standards. For example, the Ninth Revision with Clinical Modification (ICD-9-CM) was developed in the United States to provide a method of classifying morbidity data for indexing of medical records. It also includes medical case reviews, basic health statistics and other programmes. Similarly, the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision was modified by Australia and called (ICD-10-AM) to meet their local needs (Conrick, 2006). Again, ICD versions 11 and 12 were introduced to address the compatibility and interoperability issues as well as other health care challenges (Gjertsen et al., 2013).

2. SNOMED from College of American Pathologists (CAP)

SNOMED stands for Systemised Nomenclature of Medicine and was developed by the College of American Pathologists (CAP) in the USA. SNOMED CT enables the retrieval of medical information for disease management, medical research and performance analysis in order to improve quality of care. It also facilitates access to clinical information wherever it is needed across the world, whenever it is needed by any authorized user (Templin, 2006).

SNOMED is a hierarchical; multi-axial classification design system with eleven (11) axes and each of the eleven axes captures a medical concept or medical nomenclature. The SNOMED system is organised in monolithic tree structure that comprises of Topography (T) which is made up of T2800 for lung, T3200 for heart and M for organs, cells etc. there are 21 chapters on each axis that represent diseases or medical problems. For example, chapter IX represents diseases that are related to the Coronary System and in it is block 1.3 with block 110, which gives detail description of the hypertension disease and denoted as Hypertension_**IS_A** Heart disease. SNOMED also has 311,000 concepts and each concept has a unique identifier (ID) for the organised using IS_A acyclic Taxonomy (Conrick, 2006).

There are two main versions of SNOMED developed for use in electronic health records. SNOMED® RT is the next generation clinical coding with reference terminology (RT). The SNOMED RT facilitates easy transition of health care records from paper records to electronic records (Conrick, 2006). The second type is SNOMED® CT, which was created through collaboration between the CAP in USA and the United Kingdom's National Health Service (NHS) to combine SNOMED RT and Clinical Terms Version 3 of the NHS thesaurus of health care terms. The agreement created an international approach for computerizing scientific terms for physicians, nurses, and other health professionals to be used for the management of patient records and medical communication. The SNOMED CT has been adapted as the standard terminology scheme for the National Programme for IT (NPfIT) (Templin, 2006).

3. READ Codes

READ coding system is used to classify medical activities and it offers the means by which, patient's clinical records will be transferred from one GP to another GP in future. The READ system uses Linnaean classification of species of medical activity, which include illness names, operations and procedures. The aim is to allow easy transfer of information between GPs, secondary and tertiary hospitals for health care purposes. It also makes it easy for healthcare professional, clinical staff and planners to use the code (Wilcox et al., 2007).

The READ coding system is a strictly hierarchical tree structure, which comprises of branches that represent each medical classification and concepts with detail medical descriptions. It has 2 types of code descriptions namely; the 'Preferred' term, which gives a list of options and the 'Synonymous' term, which may give numerous options for one code. For example, if you enter READ code G30 the browser will give the user, a number of options such as Heart attack, Coronary thrombosis, and Cardiac rupture – MI so that the user can select their preference (Taylor, 2006).

Example of READ codes used in the EMIS software package is shown in the Figure 2.6 below: other examples of READ codes in EMIS format can be seen in (Hippisley et al., 2007)

<p>READ Clinical coding for Body Mass Index measurement</p> <p>229.. O/E is height and 22A.. O/E is weight while Body Mass Index (BMI) is 22K..%</p> <p>READ Clinical coding for Blood Pressure (BP) measurement</p> <p>246.. O/E - BP reading</p> <p>246R. Sitting diastolic blood pressure</p> <p>246Q. Sitting systolic blood pressure</p>

Figure 2.6: Examples of CVD READ codes in EMIS format (Taylor, 2006)

In spite of these advantages, the READ coding system has many limitations, which include high maintenance cost and training time. Again, the codes are updated every quarter to incorporate current changes in medical field. As mentioned for SNOMED/CR, the READ code is not internationally accepted clinical coding and therefore it is used mainly for primary care sector in UK.

The READ coding system has data compatibility and interoperability problems because it does not have inbuilt mechanism that can convert the READ code into SNOMED or ICD for statistical and other analysis (Rodrigues, 2009).

4. Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) was developed by the USA National Library of Medicine and distributed as a multi-purpose, electronic "Knowledge Sources" with associated lexical programs. The UMLS has a Metathesaurus that contains information about biomedical concepts and terms from accepted controlled standard vocabularies and classifications. The UMLS system is used for patient records, administrative health data, bibliographic and full-text databases and expert systems (Coiera, 2003).

Unlike the other clinical coding systems discussed, the structure of UMLS is based on Semantic Network, which has 134 semantic types that are linked by 54 different relationships such as 'is_a' link. The semantic structure helps to preserve the names, meanings, hierarchical contexts, attributes and inter-term relationships in the data source vocabularies (Friedman et al., 2001). However, UMLS is a repository, which collects biomedical information from different data sources and uses the Semantic Network expressions to link them. See details in Table 2.1. UMLS is an ideal model that is used for research but it cannot be used as a clinical coding tool and therefore this research will not focus on it.

5. The HL7 (Health Level Seven) - Data Representation

HL7 (Health Level Seven) is a medical coding standard for the exchange of clinical and administrative data between medical information systems. Health Level 7 is an ANSI-accredited Standards Developing Organizations (SDOs) that is used in healthcare domain such as pharmacy, medical devices, imaging or insurance (claims processing) transactions. Again, the Health Level 7 standard is designed to handle clinical and administrative data that operate on the highest application level of the ISO/OSI (International Standards Organization of Open Systems Interconnection) communications model. The HL7 application level specifies the definition of the data to be exchanged, the timing of the interchange, and the communication of certain errors to the application (Taylor, 2003).

- HL7 structure and Data Representation

The HL7 structure is based on object oriented technology and it uses the Class Model in UML to model clinical data. The primary building blocks of HL7 are the Data Models that consist of the Reference Information Model (aka RIM) and the HL7 Version-3 Meta-Model. The RIM stores the metadata of the HL7 architecture. According to Taylor, (2003), HL7 consists of 6 main classes, which are shown in Figure 2.7 and explained below:

In Figure 2.7, the Actor is the actions that are implemented and documented in the model for example, the clinical observation. The Entity is any physical thing or being that is in the problem domain under consideration such as a person, who takes part in healthcare activities. The Role is the role that the entity plays in the healthcare actions. RoleLink describes the relationship between different entities and the roles they play. Participation is known as the environment where the action takes place. Action Relationship indicates the relationship between the different actions that are performed (Jenders, 1997).

Each of the classes in Figure 2.7 is made up of attributes, which describes the characteristics of each class and their association with the RIM in the HL7 architecture. Moreover, each of these classes can also be expanded to include sub classes, where details of each of the classes are captured from the designer perspective (Taylor, 2006).

RIM Core Classes Examples	
Core Class	Examples
<i>Act</i>	<ul style="list-style-type: none"> • Clinical observation • Assessment of health condition
<i>Entity</i>	<ul style="list-style-type: none"> • Person • Chemical substance
<i>Role</i>	<ul style="list-style-type: none"> • Patient • Employee
<i>RoleLink</i>	<ul style="list-style-type: none"> • Manager has authority over Analyst (Using role link for “direct authority”).
<i>Participation</i>	<ul style="list-style-type: none"> • Surgeon • Author
<i>ActRelationship</i>	<ul style="list-style-type: none"> • Theophylline mitigates asthma (Using ActRelationship of type “mitigates”).

Figure 2.7: the RIM and the classes in the HL7 (adapted from Taylor, 2006)

The HL7 system uses SQL queries to retrieve clinical data from the CIS. The classes in the HL7 are converted into entities, as in databases to create Entity relational (ERD) or object oriented databases. SQL statements are written to retrieve relevant clinical data from the HL7 database through the RIM. For example, using the standard SQL query language and the RIM standard data model, a SQL-like query syntax, can be written as ‘SELECT last observation_value_text FROM clinical_observation WHERE abbreviation_name ISA ‘blood sugar’. This SQL-like query will allow the last observation value for blood sugar to be retrieved from the database (Jenders, 1997).

- Advantages and Disadvantages of HL7

Some of the advantages of HL7 are, it facilitates easy data exchange between heterogeneous MISs and uses object oriented technology, to implement the medical applications. The HL7 is a reference model and therefore, it can be used by any organisation to model clinical information systems. The disadvantage of HL7 is that the RIM class in the framework is changed periodically in order to incorporate modern development in the medical field therefore the core of HL7 is not stable. Another disadvantage is the sizes of the set of queries used to retrieve information are larger than the conventional relational database. It denotes that, as the data size in the database increases the query size also increases (Taylor, 2006). This leads to performance degradation and therefore query optimisation techniques must be used to enhance the performance of HL7. Again, the HL7 design structure uses the star schema, which is scalable therefore it increases the storage space requirement and creates information overload in the CIS (Jenders, 1997).

Table 2.1 Types of Clinical Coding and Data Representation Methods

Clinical terminology system	Propose	Data structure and representation (ontology)
ICD International Classification of Diseases	To collect and analyse clinical data for statistical purposes	ICD-10 is multiple-axis classification system with 21 chapters for diseases
READ CODE	For auditing clinical processes in primary care (procedure) in UK	This is a strictly hierarchical classification system
SNOMED (Systematized nomenclature of medicine)	general-purpose computer-processable terminology to represent and index all events in medical record	a hierarchical, multi-axial classification system. Terms are assigned to one of eleven modules,
(UMLS) Unified Medical Language System	Links the international terminologies into a common structure and provide translation mechanism between them	Uses Semantic Network with 134 semantic types that are linked by 54 different relationships such as 'IS_A' link;
HL7 (Health Level Seven)	standard for the exchange of clinical and administrative data Interoperability	Uses Object Oriented technique with RIM class as the core class and 6 associated classes.

2.4.2 Challenges of Data Coding and Representation in CIS

Modern CISs are designed and built using the digital image representation methods namely DiCOM, or clinical coding such as SNOMED and ICD as discussed above. These clinical coding and languages describe clinical concepts and their relationships. However, according to De Keizer et al., (2000), they are limited because they are designed with strict hierarchical structure or semantic network semantic and are developed for specific applications. The implementation of clinical data representation is limited because of the complexity of their knowledge content and the structure of the clinical data (Hoelzer et al., 2003). To address these complexities, the HL7 uses object oriented technique and UMLS uses descriptive logic with semantic network. There are also areas specific coding methods namely, Medical Subject Heading (MeSH) for United States and GALEN for Europe but they are not discussed in this research (Coiera, 2003).

These methods solve the interoperability and standardisation problems in clinical data representation but they do not address the information overload problem (Friedman, 2001).

2.5 State of the Art Solutions used for Information Overload

The clinical data representation techniques discussed in section 2.4 above do not incorporate optimisation mechanisms hence medics are compelled to retrieve the entire clinical data for diagnosis or prognosis (Taylor, 2006). Therefore the process of retrieving the entire data of instead of the relevant information needed for medical analysis, create information overload in the CIS environment. The alternative approach is to use contemporary solution or the supplementary solution which, includes optimisation techniques (Tveito and Hasvold, 2002).

2.5.1 Temporary and Contemporary Solutions for Information Overload

To address the issue of storage space complexity associated with Big Data and information overload in CIS, contemporary solution and or temporary solution are recommended. The three main recommended techniques are hardware solution, which includes distributed data centres and Cloud servers; the use of software solution that cover databases and data warehouse and the Delete Policy Solution (Ho and Tang, 2001). These solutions were mentioned in chapter 1 and their limitations are discussed below.

Hardware Solution: This approach is based on distributed data centres and Cloud servers. The hardware solution proposes that the computer hardware systems must consist of super computers (i.e. high capacity computers); distributed data centres and currently the cloud computing system. The concept is to have powerful distributed computing systems that can store the big amount of data, as well as improve the **performance** degradation and the bottle necks created in the computing network due to information overload (Velde and Van de, 2000). The disadvantage of this solution is that the failure of any of these data centres will lead to disaster in life critical applications, such as CIS. Another limitation of the hardware solution is that as the data sizes increase the computers in the networks start degrading and hence affect their performances (Ho and Tang, 2001).

Software solution uses database and distributed data warehouse, to store huge amounts of data and use SQL to retrieve data from these data sources to improve efficiency (Yoo et al., 2003). Data warehouses allow multi-dimensional gigantic information to be stored at one location using a well-defined schema (Wang et al., 1997). This approach uses data optimisation techniques to address the problems caused by poor database performance and improves the information retrieval process. However, there are still issues that relate to data warehouses and therefore, distributed data warehouse systems are implemented in order to store excessive data in the cloud computing environment. To address the data retrieval problems, the database search engines use XML with user profile and artificial intelligence techniques to reduce the delay caused by information retrieved from the Big Data sets (Park et al., 2003) but this does not address the data overload issue.

The Data Delete Policy Solution is a policy put in place by many data intensive organisations, such as banks, financial and medical institutions (Denton and Richardson, 2012). The policy states that employees must delete any information that creates information overload on the computing system. Workers are instructed to delete emails, huge files otherwise their user accounts are blocked. In medical and financial institutions this policy can have adverse effect on institution's operation. This is because mission critical application, information is cannot be deleted but it must be archived for future purposes. In the medical application, patient records and details are not supposed to be deleted, even though the size may continue to grow bigger. Therefore suitable solution must be found to address data overload in mission critical computing environment (Bertot and Choi, 2013).

From the above examples, it is inferred that the three approaches discussed cannot address the information overload and optimisation problems in CIS. Therefore alternative supplementary optimisation solutions have been recommended, which are discussed below.

2.5.2 Supplementary Optimisation Solutions used in CIS

The supplementary solution includes the use of database optimisation techniques and Clinical Client/Server Architecture (Deutsch et al., 2003). In this research, optimisation denotes the re-organisation or rewriting of the algorithm used in the CIS framework to improve its retrieval and or processing efficiency.

Another definition of software program optimization is the process of modifying a software system to make some features work more efficiently or use fewer resources. Hence a computer framework can be optimized, so that it executes faster with improved time complexity or operate with reduced space complexity, which is less memory storage spaced (Winsten and Carroll, 1996). In this research, optimization is defined as improving one or more aspects of performance of the system under consideration. The limitation and trade-off of optimization is that improvement is done at the expense of others (Fenton et al., 2003). Some of the optimisation techniques used to address the Information Overload include: optimisation of Clinical databases; the application of Best Fit Clinical Information System Architecture; and Data Aggregation.

- Optimisation of Clinical Databases

Research conducted using 59 clinical databases in Denmark concluded that to get the maximum output from the clinical databases, it must be joined together and optimised to facilitate effective data retrieval (Green, 2011). According to Sujansky et al., (1994) because of the data structure and schema of clinical data it is difficult to optimise clinical databases using the conventional database optimisation techniques. Therefore, they proposed the mapping of the Functional Relationship model (FRD) and Extended Relational Algebra (ERA). However, their evaluation and findings state that additional work must be done to achieve the required optimisation of the database in the CIS.

Johnson, (1996) proposed a generic data modelling that is used to create a database, which is better than the conversional database models and can hold large-scale clinical repository for CIS. However, this technique does not indicate how the database can be optimized to reduce information overload caused by scalability and also improve the performance of data storage. Thieke, (2007), conducted research on the storage and retrieval of Intensity-Modulated Radiotherapy (IMRT) in clinical database. It was concluded that the process is time-consuming and has many errors. Therefore, from these discussions it is established that it is difficult to invent the “best” database that can deliver an optimal CIS.

- Best Fit Clinical Information Systems Architecture:

To address the problems of information overload caused by the increase of the sizes of users' records and various applications in CIS, different types of clinical architectures have been proposed. The proposed architecture is a standard Healthcare Information System that uses DHE middleware to handle the functional aspect of the CIS and integrates different medical applications used in healthcare delivery (Massimo, 1998).

Another proposed architecture is a three-tier CIS model that uses data driven component-based system approach. In this architecture, the clinical components are designed as business objects that uses HL7, which is linked to the ERD database see section 2.2 (Hsieh et al., 2012). The third type is the Standard Clinical Client/Server Architecture, which uses object oriented technology for its implementation and was discussed in section 2.2 (Velde, 2000). These two architectures tend to increase the amount of data in the CIS because they add complex object layer onto the existing layers. For example, the application of HL7 and CORBA middleware that use object oriented technology are good but their development and implementation create extra three data layers namely the name, attribute and methods. These layers add to the existing information overload in CIS and affect its performance.

- Data Aggregation – User Profile Using XML

Data aggregation is the process of collecting data from different sources and presenting them in a summarised format for intelligent decision making or statistical analysis. A CIS application of data aggregation is to gather a lot of information about particular groups of patients based on specific variables such as age, profession, diet or life style for medical analysis. This method was applied by Park et al., (2003) to collect information about users, which was then used to retrieve personal details of users based on their known profile. The User Profile Data aggregation technique uses extendable Mark-up language (XML) technology that is applied in social media website and Internet search engine (Denton and Richardson, 2012).

2.5.3 Limitations of Existing CIS optimisation techniques

In CIS where Big Data is used, it is imperative to collect data from different sources, analyse and present them without losing their importance. Therefore the techniques used in clinical data aggregation include placing ‘relevance’ weightings on variables in the meta-data or allocating ‘reliability’ weightings as data values to facilitate easy retrieval process (Jacobson et al., 1991). Another technique used, is the Artificial Intelligence (AI), which include fuzzy logic representation method for recording of questionnaire responses. Bayesian networks, genetic algorithm and neural network algorithms are AI techniques used to study the relevance weights and patterns in systems, for predictive proposes (Friedman, 2000). It is inferred from this discussion that data aggregation techniques do not address information overload issues but are used for partial data retrieval.

It is established that the current techniques used for the capture and clinical data representation such as ICD, HL7 and UMLS do not address the information overload in CIS. The use of these methodologies in the design of CDSS does not address the information overload and optimisation issues in CIS (Fernandes et al., 2012). Hence, contemporary solution namely hardware, software and delete policy were suggested but these are only temporary solutions thus, supplementary solutions were proposed. The supplementary solution include using the Best fit CSS Architecture, database optimisation techniques and Data aggregation discussed in section 2.5.2. It was subsumed that the optimisation techniques used do not resolve the information overload issue (Winsten and Carroll, 1996).

2.6 Summary

In this chapter the role of Client Server System, as a model for designing the Architecture of CIS and role of the middleware, as decision making tool within the CDSS, were discussed. It was identified that the methods of capturing clinical data, which consist of complex multiple dimensional structure are inadequate. Again, the application of clinical data representation techniques in medical information systems, lead to the creation of Big Data and Information overload in CIS. Therefore, the existing research issues in CIS include Big Data, information overload; storage space complexity; and security issues. From this literature review, it is subsumed that further research are required on optimisation of CIS and clinical data representation to address information overload and to enhance clinical decision making, disease management and eradication.

Chapter 3: Methodology for UML -CMAUT Optimisation Framework

3.0 Introduction

This chapter 3 starts with an evaluation of the alternative conventional and contemporary Decision Making models used in the design of CDSS for CIS. This is followed by the discussion on the hypothesis, which includes the building of CMAUT prediction framework that forms the basis of this research and the methodology used to prove the hypothesis. The methodology includes the application of quantitative and statistical analysis methods, which are used in this Thesis. According to Fenton et al, (2000), Prediction Framework consists of mathematical models that work together with a set of prediction procedures for determining unknown parameters and interpreting the results. The demographic and clinical CVD data of 21,399 participants from the Health Survey for England report (HSE, 2006) were used to design and build the CMAUT prediction framework for hypertensive users. The six success criteria used to validate and verify the results of the investigation conducted using the proposed UML-CMAUT optimisation framework are presented. Finally, the discrimination and calibration methods used to confirm the hypothesis are discussed.

3.1 Clinical Decision Support System and Decision Making Models

According to Gondy and Hsinchun, (2006), in CIS architecture, the middle layer houses the CDSS and the middleware component, which is made up of algorithms that present the data to the medical decision maker. The CDSS assists in the provision of diagnosis and prognosis, which are required to evaluate diseases and recommend the requisite medical interventions.

The two main categories of decision support systems are the model for retrieving the raw medical data from the database for decision making and the model that deals with data computation and mapping the input data to the output results for decision making (Smith, 2000). The limitations of these two DSS are that they use advanced, inflexible, complex algorithms, which the designer cannot modify to meet changes in the system requirement.

Therefore, the research issues in medical decision support systems and their algorithms are contemporary fields that needs more research work (Fenton et al, 2000). Below are critical analyses of the different decisions making models used for medical decision support systems:

Analytical Decision Models for CIS: There are three types of traditional analytical healthcare decision making models: the multiple criteria decision model that deals with many decisive factors that are involved in decision making process (Fenton at al., 1997). The uncertainty decision making models deal with situation where the data that should be used for making the decision is incomplete. The third is the risk decision models where the resultant outcomes of a decision leads to risk irrespective of the decisive factors used and the uncertainty nature of the situation. The first two models use weights allocating techniques, while the utility function technique is used for the risk decision models (Sanderson et al., 2006).

3.1.1 Conventional Clinical Decision Making models

Conventional multiple criteria decision making models used in CDSS: When making decisions in health care management setting there is the need to take into consideration the multiple characteristic structures of clinical data (Fenton et al., 1997). In cases, where there is more than one criterion, the decision making requires detail analysis to find the best out of many solutions. Medical decision making involves many criterions therefore different categories of techniques are used in making multi-criterion decisions in medical decision models. According to Sanderson et al., (2006) these techniques include:

- **Utility function or Satisfaction:** also known as satisfying or utility that means setting a standard, which must be met by the individuals or functions in the decision making criterion. This is defined as the number of characteristics or attributes that every given option must meet to be considered in the solution (Duetsch et al., 1994). In econometrics, utility is used to measure the satisfaction that a customer gets after consuming a product. Utility function is also used in health decision making models (Sanderson et al. 2006).
- **Weighting:** In the weighting process, each attribute in the model is given a weight depending on the score or preference allocated to the attribute and this is compared to the required criterion to categorise the attribute as being part of the solution or not.

Weighting is used in medical applications where the medics decide, which attributes in the multi-criterion should be given higher or lower weight. In some scenarios the overall sum of the individual weights or scores is equal to 100% and this value is divided among the attributes depending upon their preferences (Fenton et al., 1997).

- **Sequential Elimination:** In this model, different options are given ranking, depending upon their importance in the problem domain. The options with the highest ranking are kept and those with the lowest are eliminated until two options are left to be analysed. This model is used when the attributes involved are not measureable and are difficult to put values on them (Sanderson et al. 2006). This model is not used in this research.

Limitations of Conventional Decision Making Models

In addition to the above, the other Decision Making Models used in medical DSS are the Games Theory, Decision Tree and Decision Graph. Games Theory is used to solve problems that have one criterion but most problems in medical application are multi-criterion (Sanderson et al., 2006). The Decision Tree model is based on the tree structure, where each branch is labelled with the probability of the nature or state of the problem actually occurring or not. The Decision graph allows the user to plot the stages involved in the decision making process from the beginning of the process to the end (Duetsch et al., 1994).

The above decision models are designed to solve problem with one criterion but most clinical applications involve solving multi-criterion problems (Sanderson et al., 2006). Again, these models are subjective and they cannot be used to solve multi-criterion problems unless they are converted into single criteria problem by applying the weighting system. This method makes the problem solving process complicated. Therefore in this research, the utility function technique is used because it is transparent and it is not subjective as compared to the sequential elimination, weighting methods and Games Theory (Duetsch et al., 1994).

3.1.2. Contemporary Clinical Decision Making models and their Limitations

Artificial Intelligent (AI) techniques such as Neural Network (NN) and Expert System (ES) are some of the techniques used in CDSS and other clinical applications (Duetsch, et al 1994).

AI is a field of science and technology that uses multi- discipline techniques to design and build intelligence machines. AI systems are designed to collect knowledge and or data from a specific clinical application and develop system that act and behave as human being. AI application software and computers are developed to operate with human intelligence and behaviour such as to reason, learn and solve problem (Laudon and Laudon, 2011). Two commonly used AI techniques in CDSS are NN and ES, their advantages and limitations are discussed below.

1. The Expert System in CDSS

Expert system (ES) is a computing system that is programmed to store knowledge and make inferences for prediction and decision making. For example, ES helps doctors diagnose diseases based on the patient's symptoms. According to Laudon and Laudon, (2011), a classic ES is made up of the following component parts:

The knowledge acquisition system is used to capture the experts' knowledge and the knowledge base is where the facts and knowledge are stored. The Inference engine is the application software that processes the knowledge and recommends the requisite action to be taken (Deutsch et al, 1994). The user interface program communicates with the end user and the explanation program in order to explain the reasons for the decision taken for the end user (Laudon and Laudon, 2011). See Figure 3.1 for the typical ES layout.

- Advantages and limitations of Expert System

The advantages of ES include the fact that it captures the expertise of the medical expert or group of experts and convert them into a computer-based information system. As an information system it is faster and more consistent than an expert and can store knowledge from multiple experts. The limitations of ES include their inability to learn new developments and changing trends in the medical field. These computer based systems are designed to solve only specific types of problems in a particular knowledge domain and therefore they have limited application. Again, many maintenance problems are associated with ES because it is designed to use the data and knowledge captured from Experts hence updating ES is a problem. The Knowledge Base, Inference Engine and rules must be changed to match the new changes and the functionalities that are required, makes it difficult to modify ES.

Therefore it is recommended that instead of modifying the existing Expert System a new system must be built for all applications (Deutsch et al, 1994).

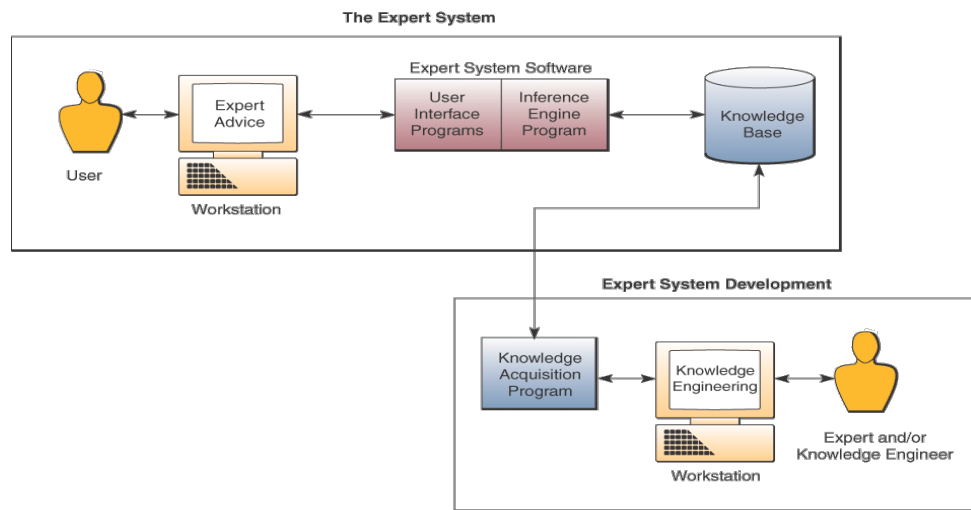


Figure 3.1 Expert System Layout and its Components Parts (Laudon and Laudon 2011)

2. Application of Neural Network (NN) in CDSS

Neural networks (NN) are computer system that functions like the human brain. Neural networks computers are modelled after the brain's mesh-like network of interconnected processing elements (aka neurons). They are composed of interconnected processors that operate in parallel and interact with each other to allow the network to learn from the data it processes (Deutsch et al, 1994). For the purpose of simulation, the Neural Net is modelled as a graph with at least three or more layers that trains the neurons until the correct output prediction pattern is identified (Laudon and Laudon, 2011).

For CDSSs that use NN application, the neurons in the input layer accept binary pattern that represents the presence (1) and absence (0) of the symptoms. The neurons in the output layer represent the diseases that the NN has been trained to recognize, where the value of 1 denotes that the diagnosis shows the present of CVD disease (Hripcsak, 1988). Alternatively, a value of 0 denotes that the patient has not got the CVD disease (Deutsch et al, 1994). For example, in the Figure 3.2, the NN system has neurons 1, 2 and 3 that represent the symptoms fever, cough and headache. The hidden layer has three hidden neurons 4, 5 and 6 that receive signals from the input neurons and then send messages to the output neurons. The output neurons 7 and 8 represent two possible CVD diseases in the final output pattern.

- Benefits and Limitations of Neural network (NN)

In the Neural Net, the knowledge acquired is stored in the structure of the network and the weights are associated with each link that connects the neurons in the adjacent layer. The neural net structure learns the weight by continuous training the neurons using the pattern of specific disease or group of diseases. The advantage of NN is that if the training of the neurons is successful, then the net is able to map its findings into diagnoses without using any disease decision making mechanism. To improve the accuracy of NN, backward prorogation is used whereby the errors from the output neurons are sent back to the input layer to modify the connection strengths in order to minimise the system error (Deutsch et al. 1994).

Although, the NN decision support systems do not rely on probabilistic approach or logical statements they have limitations in their clinical application. A study conducted by Baxt. (1991) using neural net in CDSS revealed that the outputs of NN are not precise but only indicates the likelihood of the presence or absence of acute myocardial infarction disease in a patient with chest pain. According to Baxt, (1991) after the NN have been trained and learnt the disease pattern from the input data, it performs with 92% sensitivity and specificity of 96%. These results are acceptable but Lisboa and Taktak, (2006) stated that these values are overestimated. However, NN software solutions are portable and hence can be installed on smart handheld devices that can be carried around by the medics for diagnoses.

According to Deutsch et al. (1994), the performance evaluation of NN is not accurate because they lack vigorous statistical analysis and do not include the evaluation of prevalence and Receiver Operating Characteristics/Area under the Curve (ROC/AUC) of the NN framework. Therefore in this research, the performance accuracy of the proposed CDSS CMAUT framework is verified by applying the prevalence, sensitivity, specificity, Likelihood ratio and ROC/AUC metrics. These metrics are discussed under methodology section 3.6 below.

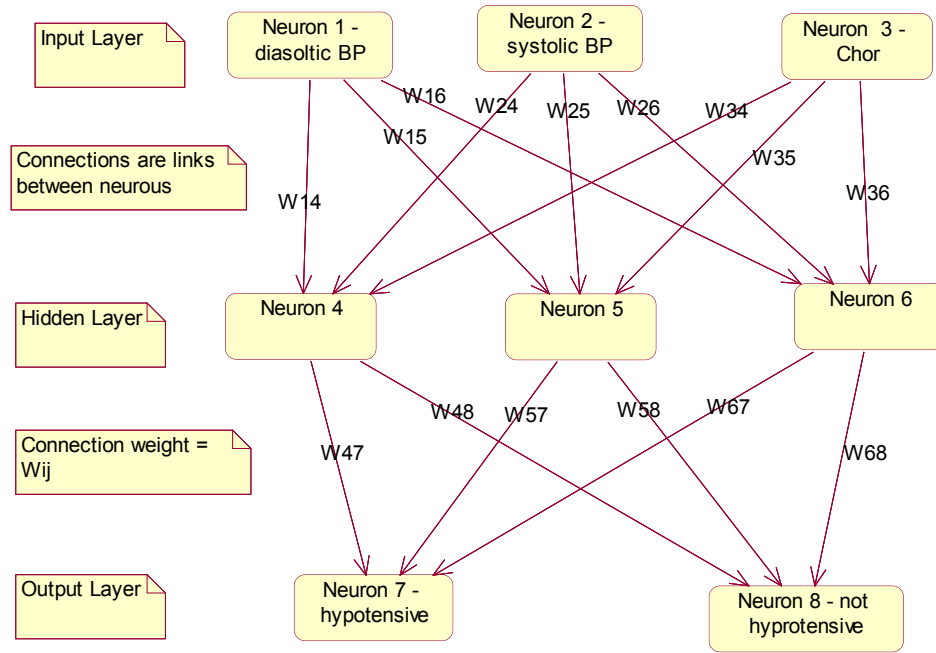


Figure 3.2 NN structure with connections, link weights and output neurons (Duetsh et al, 1994)

3.2. Research Gaps in Optimisation of Clinical Information Systems (CIS)

From the literature reviewed in chapter 2, it was established that the application of medical coding/languages as a technique used for data representation does not resolve the issue of information overload and storage space complexity in CIS (Hoelzer et al, 2003). Similarly, the use of the hardware, software and clinical database optimisation techniques are temporary solutions used to address information overload in CIS (Green, 2011). Again the application of standard CIS architecture, AI techniques and other analytical models for developing CDSS also have performance problems (Deutsch et al., 1994).

Therefore, this research adopts the clinical data re-representation approach to create a new Combinatorial Multi-Attribute Utility Theory technique that is used to build CVD optimisation Framework for CDSS. The research gap is addressed by using UML class model and combinatorial clinical components to implement prediction framework that applies CMAUT clinical decision support system for the prediction of the percentage risk of cardiovascular disease (CVD) and also reduces the information overload and space complexity in CIS.

3.2.1 Aim and Objectives:

The aim of this research is; to investigate and create an optimisation framework that uses Combinatorial Multiple Attribute Utility Theory (CMAUT) to re-represent and express clinical data in a mathematical format. Then design an algorithm that uses utility function, unit matrix or Raman transformation table (Raman et al., 1991) and LP techniques to determine the optimal amount of clinical data required for disease analysis and management. This CMAUT framework reduces the information overload, communication and space complexities encountered in CIS and improve performances.

Hypothesis: - The hypothesis to be proven is that: Clinical data can be re-represented using UML and converted into CMAUT mathematical expression with OR/AND logical connectors. The CMAUT expression can be formalised into an algorithm in a framework that can be optimised with LP technique to reduce space complexity in CIS and to predict the percentage risk of users been hypertensive as an Epidemiological Tool.

The hypothesis has been rephrased into three parts to facilitate easy referencing as follows: “1st part: Clinical data can be captured with UML and combinatorial multiple-attribute utility theory (CMAUT), which can be converted into logical mathematical expression using the data re-representation framework. 2nd part: The expression in the CMAUT framework is optimised using the linear programming (LP) technique subjected to a set of constraint matrix to reduce the data size and space complexity in CIS, 3rd part and to be used as an epidemiological tool for the prediction of Clinical absolute percentage risk and or Predictive percentage risk of user been hypertensive”.

3.2.2 Research questions:

- Can CMAUT be applied in clinical decision support system CDSS?
- How can the CMAUT be used for the representation and manipulation (aka formalisation) of clinical data in order to address the reduction of information overload in CIS?

- Which clinical applications can the CMAUT clinical data representation be used to address? Can CVD and Kidney systems be used because of their mathematical structures?
- Can CVD or Kidney related disease application domains be used to illustrate how CMAUT can be applied in CDSS to predict the percentage risk of users been hypertensive?
- How can the new technique and framework be verified to confirm the consistency of the output results and validated to establish the accuracy of the prediction framework by comparing the model performance with others using the same data?
- Will the results from the framework be benchmarked with other existing CVD risk predictor and Web CVD risk calculator?

The methodology that is used to answer these questions are analysed and discussed in section 3.4 below.

3.3 Methodology used for the Research

Different research methods have been adopted in this research because of the multi-discipline nature of the topic. Some of the fields that are covered in this research include Clinical Decision Support System, clinical data representation paradigm, risk determination, software development and combinatorial theory. The statistical and quantitative methods used are also discussed (Kremelberg, 2011).

Step 1: Systematic Analytical Review of modern CDSSs

A comprehensive systematic analytical review is conducted, to investigate and identify the state of the art methodologies used in the capture of clinical data re-representation, design and development of risk prediction models that operate with multiple attribute clinical data in CDSS. Comprehensive literature search was conducted in order to select and evaluate the three different methods used for the optimisation of CIS.

The main purpose of this investigation is to conduct in-depth research on optimisation of CIS and establish the methods and metrics used for the evaluation of risk prediction model. Again, this research process facilitates the identification of the strengths and weaknesses of the modern CIS optimisation methods. It is intended to compare and contrast these methods in order to identify, whether they address the problems specified in section 3.3.

- Risk determination and decision making models in CDSS

Risk is defined in Software Engineering as anything or any event that prevents the normal operation of software and or hardware (Fenton et al., 1997). In health care, risk is known as any activity or behaviour that endangers person's wellbeing. According to Sanderson et al., (2006), the risk decision making models used in health settings that can handle multi-attributes clinical data are Outranking, Analytical Hierarchy Process (AHP) and Multi-attribute utility theory (MAUT). These models are discussed in Chapter 4.

In medical application, the three types of risks are absolute, predictive and relative risk (Sanderson et al., 2006). In this research absolute risk is defined as the probability of a person having a disease based on their current multiple attribute values, which are measured during the time of their medical examination. Predictive risk is the possibility of a person developing a disease such as CVD over a specified time-period for example 5 or 10 years based on their present circumstances and the values of their measured multiple attributes. Relative Risk is the comparison of risk between two different groups of people under consideration (Panagiotakos and Stavrinos, 2006).

This research focuses on clinical Absolute Percentage Risk (APR), which is associated with diagnosis of a user based on the clinical conditions when during medical examination and the recorded multiple risk attributes values at the time of examination. Predictive percentage risk (PPR), on the other hand is, the percentage possibility that a participant undergoing medical examination would develop the CVD disease over a specified period of 5 or 10 years based on their present circumstances and the values of the measured multiple risk attributes.

Step 2: Development and Creation of a new CMAUT optimisation framework

A feasibility study was conducted to determine whether the existing clinical data representation can be extended using CMAUT and mathematical representations.

Then comparative analysis of different data representation methods, which include descriptive modelling, data formalism with First Order Logic and mathematical representation with CMAUT were performed. Based on the findings from the comparative analysis, a framework was developed incorporating the identified unique characteristics needed to address the problems associated with the optimisation of CISs.

3.3.1. Domain scenarios for modelling the CMAUT Prediction Framework

The two domain scenarios that will be used to develop and verify the framework are: Kidney and Heart related diseases: -

Scenario 1:- The first case study focuses on kidney related diseases, which involves the two kidneys that can perform the function of each other. In the human system, each kidney can replace or substitute the other during their operation. According to Guyton and Hall (2006), kidney related diseases are classified into two groups, which are acute renal failures, where the kidneys stop working entirely or may stop for a period but may recover back to its normal function. Secondly, is the chronic renal failure where is progressive loss of function with more nephrons that gradually decreases overall kidney function. This research focuses on Interarenal acute renal failure, which is the result of abnormal behaviour of the kidney itself.

Scenario 2: The second case study is based on heart related diseases, which are heart failure, hypotension, hypertension and angina. This research focuses on hypertension (aka high blood pressure) and hypotension (aka low blood pressure). The hypotension and hypertension diseases affect the three primary organs namely the heart, kidney and Ant Diuretic Hormone (AHD) in the brain, which complements each other in their operations. Figure 3.3 shows the cardiovascular system and the main organs namely Kidneys, lungs and brain, which are used in this research. These diseases were selected because they are complex and they affect more than two organs, furthermore there are a lot of data on them that can be used for validation of the framework (Guyton and Hall, 2006).

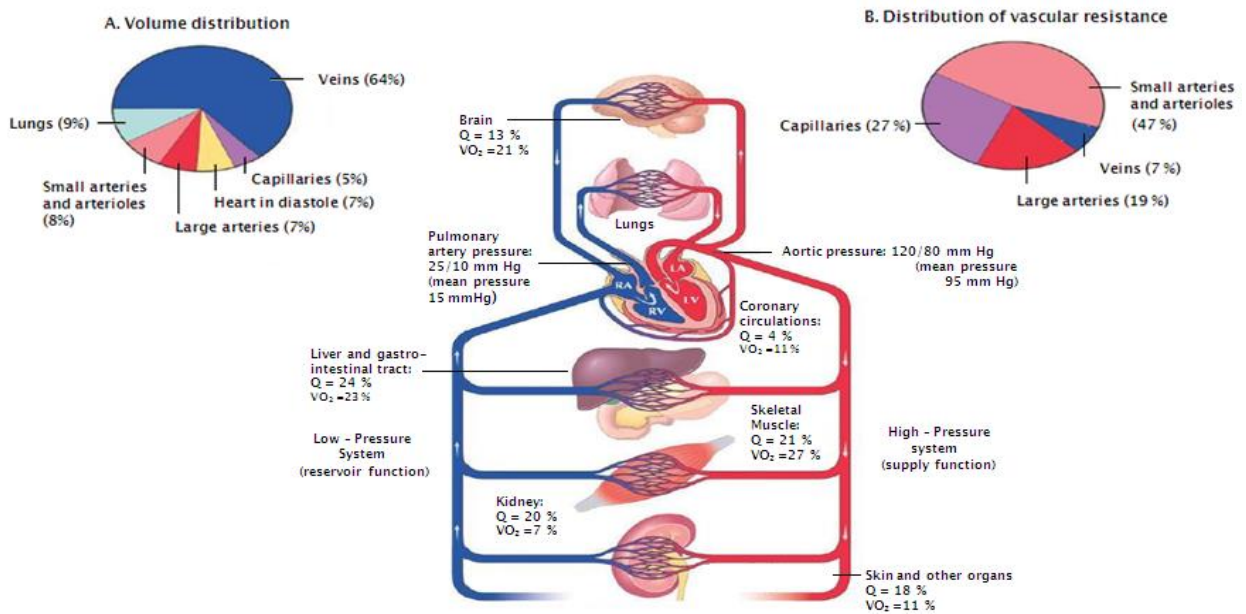


Figure 3.3 the cardiovascular system and the main organs, (Guyton and Hall, 2006)

3.4 Data Collection and Analysis

Stage 1:- Clinical data for the heart disease specified in the scenario was collected from medical literature, which include (Guyton and Hall, 2006) and (Templin, 2006). According to Templin, (2006), clinical data have acceptable measurable parameters used to analyse and manage CVD diseases. The minimum and maximum boundary values of CVD risk parameters are recorded in Table 3.1. These clinical data are used in the algorithm to determine the optimal amount of data required for CVD disease analysis. These data values are used to test the soundness of the CMAUT clinical data re-representation and to validate and verify the optimisation algorithm in the CMAUT framework.

Stage 2:- The anonymous demographic data provides the description of each participant's age, height, weight and sex but their names and other personal details are excluded. The clinical data describes the participant's disease and the measured attribute values required to analyse and management heart disease. The results of the APR and PPR values from the CMAUT framework will be benchmarked against the values of from the Internet based CVD Calculators that use Framingham algorithm and comparative analysis carried out (Sheridan et al., 2003).

The data set used to design and build the CVD optimisation risk prediction framework, was the primary raw data from the 21,399 individual records of people who participated in the 2006 cardiovascular disease (CVD) survey conducted by Health Survey for England (HSE). The data from the HSE, (2006) report was filtered to include the relevant data required for the designing and modelling of the optimisation framework, which is further discussed below.

3.4.1 Data for modelling and simulation of CMAUT CVD Framework (HSE, 2006)

The demographic details and clinical data used in this research are from Health Survey for England (HSE) report undertaken by University College London (UCL) in 2006 on cardiovascular disease and risk factors in England (HSE, 2006). The full survey report with the detailed data of all the 21,399 individual participant records was given out for this research work with the permission of the authority of National Centre for Social Research for UK Data Archive. The original data, which was in SPSS, was exported into MS Excel spreadsheet for analysis. According to Craig et al.,(2006a), the methodology used for the data collection, includes measuring each variable three times, compute the mean value and crosschecking the values with the specialists. Other techniques used to ensure that the individual levels of information collected are accurate are discussed in Craig et al., (2008). The report also states that two samples were used namely the core sample, which are adults of over 16 years and boost sample of children between the aged 2 to 15 years.

- HSE Data used for designing the CMAUT Framework

From the HSE, (2006) report in SPSS, the demography and clinical data of all participants who provided complete detailed data were first filtered out. Therefore out of the full data provided only the relevant demography data was included in the preliminary filtering process. The initial stage of the data analysis was to filter out and eliminate all incomplete data from the original survey report that comprise of 21,399 individual records. Only participants whose full clinical data were measured and recorded were selected (see Appendix 3.1C for the HSE06ai full data in Excel Electronic format).

The First stage of the filtering process for relevant data from the 21,399 individual records was carried out as follows: it was identified that only 9194 participants had full clinical and demographical data that were relevant to the development of prediction models. Therefore

these participants' data sets were filtered to ensure that all clinical data attributes required, were included. The first data sheet is made up of the core adult sample, which are over 16 years old and some of the boost children between the 2 and 15 years old, who refused to have the HDL test. (See Appendix 3.2C for the HSE06ai filtered data of 9194 records in Electronic format).

The second stage of the filtering process was to eliminate all children, under the age of 16 years from the data sheet. This is because the Framingham's equations for CVD risk analysis and prognosis have been designed for adults of over 20 years old (Sheridan et al, 2003). The result of the filtering process gave the list of adult participants who are over 16 years old as shown in the Excel spread sheet Appendix 3.3. The total number of participants in the records is 4316. (See Appendix 3.3C CVD data set of participants of over 16 years old for Model I with 4316 records in Electronic format).

The third stage of the filtering process was the elimination of non-directly related clinical CVD variables or attribute values, such as household size, income among others. Therefore only the measurable and non-measurable clinical CVD risk prediction variables and attributes were kept. Again, research conducted on Internet CVD risk calculators, which are discussed in chapter 8 and used for benchmarking the results from the CMAUT framework, are designed for people between the ages of 32 and 72 years (Chuang et al, 2007). Owing to this constraint, the (HSE, 2006) was filtered to exclude all participants who are less than 30 years. Therefore the final individual records used in this research to test the CMAUT framework models 1 and 2 are 3645 participants. (Appendix 3.4C is the CVD data set of the 3645 participants over 30 years old records used for Model 2 and simulation in Electronic format).

3.4.2 Design and Implementation of the CMAUT Framework;

The two prediction models built in this research were designed and implemented using the Variables in the Equation from SPSS logistic regression. To build the CVD risk prediction framework, the binary logistic regression method in SPSS was used and beta coefficient determined for the formation of the structural equation model. This is further discussed in chapter 5, where the diagnosis framework is modelled and in Chapter 7 where the prognosis framework is discussed (Kremelberg, 2011).

The first CMAUT model 1 was built using “Variables in the Equation” obtained from analysing the data of 4316 adults over 16 years who provided full clinical data see Appendix 3.4 (HSE, 2006). This is because the first core sample used for the CVD survey was based on adults who are over 16 years old. Again medical literature on clinical analysis and management of CVD include adults who are over 16 years old age group (Craig et al., 2006a) and (Craig et al., 2008).

The second CMAUT model 2 was built using “Variables in the Equation” obtained from statistical analyses of the HSE 2006 data of 3645 participants who are over 30 years with full clinical data. The APR and PPR results from the CMAUT framework are compared with the percentage predictive risk (PPR) results from other Framingham equation and web calculators. This is discussed in chapter 8.

- **Rationale for the selection of CVD risk parameters for CMAUT Framework:**

According to Anderson et al., (1991), the probability of a person developing CVD disease over a period of time is computed using the measurable and non-measurable risk factors. The Framingham CVD algorithm uses mainly non-measurable risk factors such as age, sex, HBP, ethnic origin, Diabetic and Smoking to predict the percentage risk (Wilson et al. 1998). This is confirmed in the CHD Web Risk Calculators research conducted by Sheridan et al., (2003).

However, this research subsumes that the development in medical science requires that precise and accurate risk factors must be used for the prediction of CVD disease. The risk parameters used for this research are the recommended CVD factors in (NICE, 2006) report. These risk parameters are a combination of measurable and non-measurable risk factors.

The selected risk factors for the framework in Table 3.1 and Table 5.1 are explained below.

1. The selected demographical data includes each participant's series number, age sex, ethinda (i.e. ethnic origin);
2. The main clinical data used include HB, HDL, BMI, TC, HDL BPH, BPL, MAP, ECG/LVC and CVD, Diabetic and Smoking.

The parameters were selected to maintain consistency, when benchmarking the PPR results from CMAUT models with the Framingham equations and Web Risk Calculators results.

Secondly, in CIS, prediction models are verified using discrimination and calibration techniques therefore these risk parameters were selected to ensure that there are similarities between the risk factors used by the different existing prediction models to avoid prejudices.

Table 3.1: The measurable Clinical data used in the research (source: HSE, 2006)

Attributes Used in work	Medical name	READ code	Baseline value	Min value in HSE	Max value in HSE
OmpulvalHB	Heart Pulse Rate (Beats Per Minute)	NA	50 BPM	20.50	136.50
BMI	Body mass index	22K..%	25.5 kg/m2	13.20	49.66
OmsysvalBPH	Sitting systolic blood pressure	246R.	140 mmHg	84.00	225.00
OmdiaBPL	Sitting diastolic blood pressure	246Q.	90 mmHg	39.00	117.50
Hdlval1HDL	High-density lipoprotein (HDL) cholesterol	NA	1.2 mmol/L	0.50	4.40
OmmapvalMAP	Mean Atrial Pressure	NA	100 mmHg	58.50	140.00
CholvalTotalChol estrol	Total cholesterol measurement	44PH.	5.5 mmol/L	2.10	11.40

- **Packages used for the development and implementation of CMAUT Framework**

The CVD clinical data was analysed using the Statistical SPSS Software package version 15. The SPSS was used to determine the “Variables in the Equation” that is required for the construction of the structural equation model and implementation of the CMAUT framework. The CMAUT Prognosis and Diagnosis prototype frameworks were built using the MATLAB (Mathematical Laboratory) software version 2007b. The MATLAB software has toolboxes such as bioinformatics, optimisation and statistical for analyses. MATLAB was also used to generate tables, reports, charts and conduct complex statistical analyses. The algorithm in the frameworks was designed with utility function and Unit Matrix that use the LP technique in MATLAB software. The SPSS, Excel and the MATLAB statistical tool were used to evaluate the hypothesis that: “the clinical data representation can be re-represented using CMAUT expression in mathematical formalism that serves as input to an optimisation algorithm”.

The Rational Rose software was used to model the organs in the disease domain. The class diagram was used to model the organs and the relationship between them. The substitutable organs in the domain were modelled as sub-classes and the complementary organs were modelled as direct associations. The multiple attributes of each organ were represented as the clinical data and the functions of the organ denoted as methods in the class diagram.

3.5. Application of Statistical and Quantitative Methods

In this research, a combination of statistical and quantitative methods was used to design, implement and evaluate the CMAUT framework. To achieve the research objectives, the following metrics and methods were used. They are Clinical Absolute Percentage Risks (APR) and Predictive Percentage Risks (PPR); computation of the Prevalence and Kappa values; determination of space complexity, computation of the data sizes before and after optimisation using P-value; calculation of sensitivity and selectivity; determination of the accuracy of prediction models using AUC/ROC and Likelihood Ratio.

3.5.1. Statistical Methods

Statistical methods were used to compute the correlation and relationship between the different variables in the CVD data set from the HSE, (2006) report. The independent variables used to predict the output percentage risk, were based on risk factors selected from existing CVD risk prediction models. The existing prediction models are Framingham equations and Web based CVD Risk calculators, discussed in Chapter 8. The independent variables used in the CMAUT framework are HB, HDL, BMI, TC, HDL BPH, BPL, ECG/LVC and CVD, Diabetic and Smoking. The dependent variables are the output from the optimisation framework which are the Clinical Absolute Percentage Risk (APR) and Predictive Percentage Risk (PPR) (Kremelberg, 2011).

To prove the data size reduction part of the hypothesis, the P-value was used to determine the difference between the data sizes before optimisation and after optimisation with CMAUT framework. The statistical significance analysis uses the independent pair samples T-test technique, which are the difference between the two groups of data sets before and after optimisation (Kremelberg, 2011). The pair T-test, was conducted using a random sample data set of 402 participants, which is approximately 10% of the total number of 3645 participants who submitted full CVD data and were selected from the (HSE, 2006) report.

Functional performance measures the storage space change in the prediction framework before and after the framework has been enhanced (Fenton et al., 2007). The space complexity is used to validate the amount of data needed for management of the CVD before and after optimisation using the CMAUT framework.

The aim of this research is to determine if the memory space and space complexity have increases or reduces before and after the application of the CVD predictive risk model in CIS. The evaluation is conducted using the Big O notation, which is discussed in chapter 6 and 9. This is followed by the statistical analysis of the data sizes before and after optimisation using the CMAUT framework.

3.5.2 Quantitative Methods

1. Cohen Kappa

The Cohen Kappa is used to measure the degree of agreement between the predicted results from the framework and the GP suggested diagnosis. Kappa statistic is defined as the measure of a system based on the degree of reliability of the agreement between two predictors. According to Viera, (2005), Kappa is a method of the determining the degree of agreement between two predictors using quantitative measure. Kappa is an important validation because it reduces the possibility of chances in the prediction. In this research, kappa is use to determine the degree of agreement of the actual percentage risk results from the HSE, (2006) survey and the risk results for the two CMAUT Framework models.

The Table 3.2 is built using the output risk results from the two predictors. The kappa k value is computed using the formula from (Cunningham et al., 2009) and (Viera, 2005). These formulae (3.1) are explained and used in chapter 5 and 7.

Table 3.2: Table for computation of Cohen Kappa statistic:

		Predictor 1 CMAUTF		
		Yes	No	
Predictor 2 Actual	Yes	a	b	$a + b = m_1$
	No	c	d	$c + d = m_0$
Totals		$a + c = n_1$	$b + d = n_0$	N

The formula is $k = \frac{P_o - P_e}{1 - P_e}$ where

$$P_e = [(n_1/n) * (m_1/n) + ((n_0/n) * (m_0/n))] \text{ and } P_o = [(a + d)/N] \quad (3.1)$$

2. Prevalence

The computed Prevalence value indicates the presence of hypertension disease in the population of participants who took part in the HSE, 2006 survey. Prevalence is used by

medics, health providers and epidemiologists to estimate how common the disease under consideration is found in the population and in an area at a particular time (Campbell, 2007). To compute the prevalence value the generic mathematical formula is: $\text{Prevalence} = a / (a+b)$ and for this research the formula used is:

$$\text{Actual Prevalence} = GP \text{ bpyes} / (GP \text{ bpyes} + GP \text{ bpno}) \quad (3.2)$$

In these equations, ‘a’ or ‘GPbpyes’ is the number of participants in the surveyed population who were diagnosed as been hypertensive during the survey. The ‘b’ or ‘GPbpno’ represents the people in the surveyed population who were not hypertensive during the survey. This equation is used in chapters 5 and 7.

the NICE, (2006) threshold is used to determine, whether a participant has hypertension (i.e. BPyes) or does not have hypertension (ie BPno). According to NICE (2006), the threshold of the percentage risk of CVD must be less than 20%. Therefore any person whose percentage risk value is higher or equal to 20% must be diagnosed as hypertension YES while any person with risk value of less than 20% must be declared as hypertension NO. For the purpose of diagnosis and prognosis, the GP identified hypertension are marked YES (1) and non-hypertension marked NO (0). These values are compared to the results from the Prediction Model based on the NICE 20% recommendation to avoid any prejudices.

3. Sensitivity and Specificity:

According to Campbell et al. (2007), the accurate methods for the validation of medical risk prediction models are the use of sensitivity, specificity analysis and its associated receiver operating characteristic with the area under the curve (ROC/AUC) method. Sensitivity is also known as True Positive Rate (TPR). Sensitivity is expressed in percentage and it is defined as the probability that when the test results of participant is positive then it indicates that the participant has the CVD disease. Specificity is known as False Positive Rate (FPR). Specificity is expressed in percentage and it shows the probability that when the test result of participant is negative then the participant does not have the CVD disease.

The formulae for calculating the Sensitivity and Specificity or TPR and FPR are as follow:

$$\text{Sensitivity} = \text{TPR} = \text{TP} / (\text{TP} + \text{FN}); \text{ and the Specificity} = \text{FPR} = \text{TN} / (\text{FP} + \text{TN}); \quad (3.3)$$

In the formula (3.3), True Positive (TP) is the fraction of the population who were accessed and classified as having the hypertensive disease. The False Negative (FN) is the fraction of the population, who were accessed and classified as not having the hypertension disease. However, there is a portion of the population who are without the hypertension disease and have been classified as negative (TN) also known as True Negative fraction. While others who are without the hypertension disease but are classified as False Positive fraction also known as positive (FP).

An MS Excel procedure was used to determine the TPR and FPR of each Risk prediction model, which is explained in Figure 3.4 below. The threshold for the CVD percentage risk is 20% (NICE, 2003). Therefore the calculation of the TPR and FPR is done by comparing the individual percentage risk value from the CMAUT framework with the recommended 20%.

1. Copy the percentage risk valves from the Risk model into an Excel Spread-sheet
2. Create a column that records the values that exceed the NICE risk of 20% and name it as “Exceed”. In this column, if any value is greater than the 20% threshold it is entered as 1 else input 0.
3. Create another column called “Non Exceed”. In it enter 0 if the percentage risk is less than the 20% threshold else enter 1.
4. Using the formulae (3.3) below calculate the TPR and FPR:
 - $TPR = \text{Sum (Exceed 2: All Exceed)} / \text{Sum (Exceed 1: All Exceed)}$
 - $FPR = \text{Sum (Non Exceed 2: All Non-Exceed)} / \text{Sum (Exceed 1: All Non- Exceed)}$
5. The results are depicted in model **Tables**

Figure 3.4: The procedure for computation of Sensitivity (TPR) and Specificity (FPR)

The sensitivity and specificity values from the CMAUT framework and CVD risk calculators were computed using the procedure above. In this research, the sensitivity and specificity values give an indication of how the probability of having a disease can change when there is a positive or a negative outcome.

4. Positive and Negative Likelihood Ratio (LR+) (LR-)

The sensitivity and specificity values were used to compute the probability of a participant having the CVD disease when there is a change in their LR results from positive to negative.

Likelihood ratio is used to determine the optimal risk cut-off value of the CVD Risk prediction models based on the NICE recommended value of 20%. The Positive Likelihood Ratio (LR+) and Negative Likelihood Ratio (LR-) values are calculated using the sensitivity and specificity results as follows.

$$\begin{aligned} \text{Likelihood ratio positive (LR}^+ \text{)} &= \text{Sensitivity} / (1 - \text{specificity}); \\ \text{Likelihood ratio negative (LR}^- \text{)} &= (1 - \text{Sensitivity}) / \text{specificity}; \end{aligned} \quad (3.4)$$

The equations in (3.4) are expressed as Positive Likelihood ratio that is $LR^+ = (TPR/1-TNR)$ while Negative Likelihood ratio is $LR^- = (1-TPR/TNR)$. In graphical format all the Positive and Negative Likelihood ratio values are plotted on the Y-axis and the participant identification (PIND) of each participant is plotted on the X-axis. The interception of the Positive and Negative Likelihood Ratio curves are compared against the criterion value. When the values of (LR+) are higher, they are considered as better while lower values of (LR-) are better (Campbell 2007).

5. Receiver Operating Characteristics and the Area under the Curve (ROC/AUC)

The Delong Approximate Trapezoidal method was used to determine the area under the curve. In this technique the first step is to plot the TPR against the FPR values as a curve for each model. Then the best curve fit method is used to find the equation that best fit the curve. This is followed by applying the Trapezoidal rule where the area under the curve is split into a number of trapeziums and their areas calculated. Finally the summation technique is used to determine the approximate value of the area under the given curve. The Area under the curve AUC is computed using the Trapezoidal formulae:

$$((\text{Sensitivity values} + \text{Specificity values})/2) \text{ or } ((\text{TPR values} + \text{FPR values})/2). \quad (3.5)$$

In this expression, the diagonal reference line gives a value of 0.5, which is half of the square of the area under consideration (Cui, 2009).

3.6 Validation and Verification of the CMAUT-CVD CIS Framework

In this research, the CMAU framework was evaluated using medical validation and verification techniques. In this context, verification means the process of confirming that the output of CMAUT framework is viable.

Validation means the degree of accuracy of the output PPR values from the framework when they are compared with the HSE, (2006) CVD results. According to Sanderson, (2006), in medical models, validation is used to establish the relationship between the results from the predictive framework and the theoretical concepts.

The method used for the validation of CVD epidemiological model is calibration, which measures how the PPR values agree with the actual results suggested by the GP (Panagiotakos and Stavrinou, 2006). The metrics used for calibration of predictive model are Kappa, which measures the degree of agreement between predicted results from the framework and GP's suggested diagnosis. The second metric is Prevalence, which depicts the presence of hypertension disease in the population of participants who took part in the HSE 2006 survey (Campbell, 2007). The third metrics is the computation of the sensitivity and specificity values of each of the prediction models using the NICE threshold value of 20%. This is used to benchmark the PPR values from the CVD -CMAUT framework against the results from the Internet CHD calculators and Framingham CHD algorithms.

Verification also means if the output risk values from the prediction models are consistent with the observed results from the GP predicted CVD YES or NO as well as other Web based CVD calculators (Sanderson, 2006). For epidemiological models, verification is carried out using discrimination method (Panagiotakos and Stavrinou, 2006). Discrimination is the ability of prediction model to assign higher probability values to participants who have the CVD disease as compared to those who do not have the CVD disease. This metric is used to quantify the discriminatory ability of models are C-statistic and receiver operating characteristics (ROC). In this research, ROC/AUC deals with the degree of the accuracy of results from the prediction models. The discrimination ability of a model is also determined using the Likelihood Ratio, which indicates how changes in the test results of a participant will have effect on the probability the participant having the disease or not.

Incidence is another form of verification metric and it is defined as the measure of risk of people in a given population that will develop a particular disease over a specific period of time (Campbell, 2007). This incidence measurement is expressed as occurrence rate of the disease happening over a period of time such as 5 or 10 years. This validation method requires that a follow-up survey is conducted.

At the time of writing this thesis the HSE, (2006) the follow up survey report had not been released therefore, incidence rate is not included in this research.

3.7 Summary

In this chapter, the CDSS was discussed. It was established that MAUT would be suitable for modelling CDSS to facilitate the capture of the appropriate amount of data required for clinical analysis. From the literature review conducted it is proposed that the extension of data re-representation technique with a hybrid of Combinatorial theory and MAUT will resolve the information overload issue and create a new CDSS modelling paradigm in CIS. Therefore in section 3.2, a research gap with aims and objectives was established.

To achieve the objectives, a set of methodologies were discussed in this chapter, which include the design and implementation of the proposed CMAUT framework for hypertension user. Again, six success criteria have been identified for the verification and validation of CVD prediction models. The PPR results from the CMAUT framework would be benchmarked against existing CVD risk prediction models such as Framingham algorithms and Web based Risk calculators. Finally, the calibration and discrimination ability of the proposed diagnosis and prognosis CMAUT frameworks as well as the other CVD risk prediction models are discussed in evaluation chapter 9.

Chapter 4: Optimization and Clinical Data Re-representation in CIS

4.0. Introduction

In this Chapter, the two clinical data re-representation techniques, ERD/FOL proposed by De Keizer and Abu-Hanna, (2000) and Entity Attribute Value /Class Relationship (EAV/CR) are evaluated. This is followed by a detailed review of the different decision making models used in medical and Clinical Decisions Support Systems (CDSS) namely Outranking, Analytical Hierarchy Process (AHP) and Multi-attribute utility theory (MAUT).

The Chapter finishes with an explanation of how the proposed UML- CMAUT framework is used to capture clinical data that can be optimised to reduce information overload. This is followed by using examples to illustrate how the framework models kidney and CVD disease using UML class diagram. It explains how the multiple attributes represented in classes and applied to write CMAUT mathematical expression using utility unit and logical connectors.

4.1 Modern Clinical Data Re-representation Methods

The two modern clinical Data Re-representation techniques are FOL/ERD and the EAV with object oriented enhancement (EAV/CR). These clinical data re-representation applications are based on the Data Re-representation method proposed by Haimowitz et al., (1988). In the article "Representing Medical Knowledge in Terminological Language is Difficult", Haimowitz et al., (1998) presented the concept of capturing and re-representing medical data structure with appropriate format in order to facilitate cost effective modelling and design of medical framework.

4.1.1 First Order Logic/Entity Relationship Diagram (FOL/ERD)

Data re-representation technique is a computing problem that has been researched on and adapted by De Keizer and Abu-Hanna, (2000) to address the issue of clinical data standardisation and compatibility. De Keizer et al., (2000) analysed the ontology and structure of different medical languages and highlighted their limitations. Then they proposed a new clinical data re-representation that uses Entity Relationship Diagram (ERD) formalism to express clinical concepts and the relationship between them.

To complement the conceptual model, De Keizer and Abu-Hanna, (2000) used first order logic (FOL) as an expressive instrument and formal specification (i.e. mathematical format) to avoid ambiguity. The two component parts of FOL/ERD are the ERD, which is the conceptual model while the FOL is used as the retrieval mechanism. Figure 4.1 shows the data re-representation format for ICD, which is different from the original mono-axe hierarchy structure of ICD. This approach increases the usability of CIS and confirms the fact that using only descriptive logics is not enough for modelling CIS (Ceustersa et al., 2003).

- The FOL/ERD Conceptual Model:

The conceptual model in Figure 4.1 indicates the main data representation where the medical concepts are represented as squares while the attribute of each of the concepts are shown as oval. The relationship between the concepts is shown using arrows with the nomenclature such as “Is_a”. Other associations between different concepts are Extension and the Designate code. Again, when a concept is a subordinate concept of the main concept the relationship between them is represented as self-reflective relationship and it is termed “Direct_Is_a”. This paradigm of Data Re-representations is discussed in details in (De Keizer et al., 1999), (De Keizer et al., 2000a) and (De Keizer et al., 2000b).

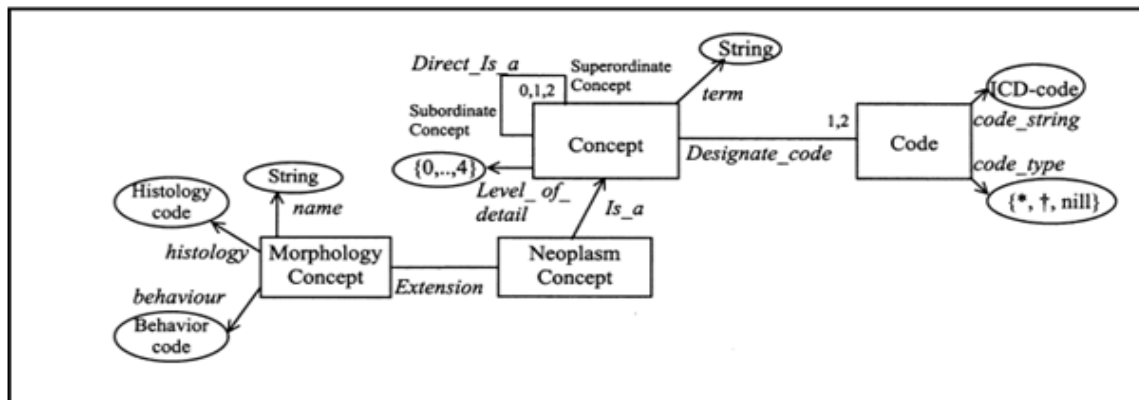


Figure 4.1: Clinical Data Re-representation of ICD using FOL/ERD (De Keizer et al, 2000a)

The FOL is based on relational calculus, which is used to capture the relationship between the different concepts in Data Re-representation framework as shown in Figure 4.1. This is followed by algebraic calculus that medics can use to retrieve data for decision making. In this decision making technique, the request in the disease domain is expressed in formal specification using first order logic (FOL) and presented by algebraic calculus in a mathematical format to avoid ambiguity (De Keizer et al., 2000a).

For example, each concept and its relationship as shown in Figure 4.1 as 0 to 4 is written in FOL as $\forall c \in \text{Concept } \text{Level_of_detail}(c) \leq 4$ as in the Figure 4.2. This expression cannot be written by non-specialist; therefore the framework has not been implemented in real life. In the ERD, relational algebra is used to write the data to be retrieved for decision making in formal specification however, this requires complex query expression. More information can be seen in De Keizer et al., (2000b).

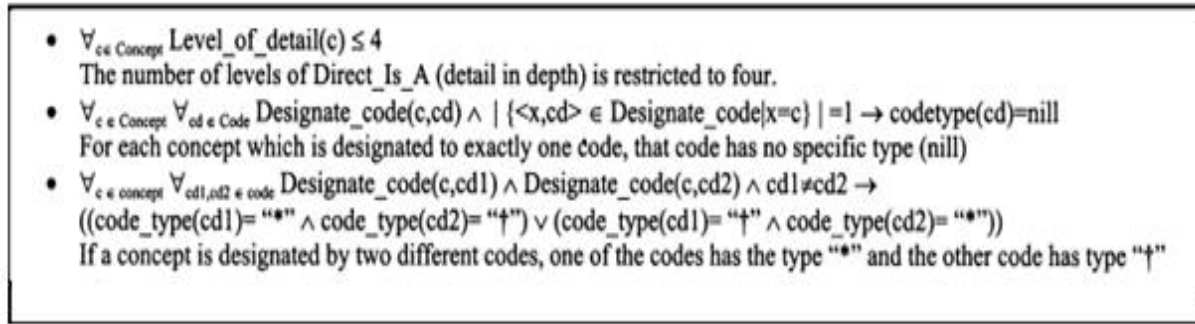


Figure 4.2: Clinical Data Re-representation of ICD using FOL/ERD formalism with relational calculus (De Keizer et al., 2000a)

Limitations and Challenges of FOL/ERD are:

1. The FOL/ERD clinical Data re-representation method is a user friend conceptual framework as shown in Figure 4.1 because it follows the basic notations and models used in Data Flow Diagram, ERD and oriented objected database. For example, in the conceptual schema, medical concepts and codes are represented as squares while attributes are indicated as oval that are attached the squares. Arrows indicate the flow and association between the medical concepts and codes. Although, this framework is helpful, it has many limitations that create anomalies and ambiguity.
2. Although the FOL was introduced to address the anomalies and ambiguity issues, it requires excellent knowledge of Formal Methods to write them (Charatan and Kans 2003). Again, this procedure must be incorporated into the Clinical Decision Support System to make it effective. The dynamic nature of clinical data means the FOL based system must be written for real-time application.

Lastly, the FOL/ERD has not been implemented for real life application therefore its performance cannot be established (Abu-Hanna et al., 2004).

Abu-Hanna et al. (2004) designed and implemented a prototype FOL/ERD system using Protégé. However, the prototype does not have an optimisation system therefore the challenges of Information Overload using this framework still remain unresolved.

4.1.2 Entity Attribute Value/Class Relation (EAV/CR) and ER Model:

To address the issues of big data in clinical environment and facilitate the efficient storage of multidimensional clinical data, Entity Attribute Value was introduced by Nadkarni, et al., (2001). This concept was enhanced by incorporating object oriented technology, which is called Entity Attribute Value/Class Relation–EAV/CR. The main advantage of EAV/CR is to facilitate the storage of clinical data and its relationship with other classes. The EAV/CR captures the values and attributes of the classes, with respect to time (Nadkarni and Brandt., 1998). For example a patient can be seen by different doctors many times in a day and different clinical data values measured and recorded in the EAV database simultaneously. The EAV allows the storage of this data without overriding the previous data.

The Entity–Attribute-Value Model is a standard data model used in the electronic healthcare because of its unique multiple attribute characteristic (Anhøj, 2003). In the EAV model, each row in the table represents triple components, which are the entity, an attribute and an attribute value. For instance, in Table 4.1 below the entity “patient” has the attribute “DOB” a value of which would be text “30-08-85”.

In contrast to EAV, the Entity Relation Diagram (aka ERD) is used for the design of conventional databases (Silberschatz, 2001). ER database is a two dimensional container in which data is stored in columns and rows. In an ER model, entity represents a discrete object or concept and the relationship between entities are known as association. ER Model is a conceptual schema that uses relational model where an entity is implemented as table and the attributes of the object are implemented as columns in the Table. The relationships between the different entities are shown with arrows and implemented either as separate tables or foreign key columns in an entity table in Table 4.2.

Table 4.1: Conventional ERD schema - Relational Database Design

Patient ID	Name	DOB
1	William Henry	30-08-85
2	Peter Jackson	02-12-85

Table 4.2: Entity Attribute Value (EAV) Database Design

Patient ID	Attribute	value
1	Name	William Henry
1	DOB	30-08-85
2	Name	Peter Jackson
2	DOB	02-12-84

- Querying Clinical Data in EAV:

From point of view database, querying EAV database is similar to querying the conventional ERD database because they both use Structured Query Language (SQL). For example, querying Table 4.1 to retrieve information about a patient whose name starts with William and who was born before 1985 is simple as shown in Figure 4.3. However, the EAV database query requires more predicates because both the attribute column need to be specified. In the EAV database query, the Table 4.2 needs to be self-joined in order to return the name of the patient as shown in Figure 4.4 below. Therefore generating queries for EAV model is more complex as compared to conventional database (Anhøj, 2003).

```
SELECT *
FROM Table 4.1
WHERE Name LIKE 'William%'
AND DOB < '1985';
```

Figure 4.3: SQL for data retrieval from Conventional ERD Table

To retrieve the same result from the EAV table requires more complex SQL:

```
SELECT d1.patient ID AS patient ID,
       d1.value AS name,
       D2.value AS DOB
FROM Table 4.2 AS d1 INNER JOIN Table2 AS d2
USING (patient ID)
WHERE d1.attribute='name'
AND    d1.value LIKE 'William%'
AND    d2.attribute ='DOB'
AND    d2.value< '1985';
```

Figure 4.4: SQL for data retrieval from EAV Table

- Entity Attribute Value/Class Relations using Object Oriented Technology:

The object-oriented method for modelling EAV is known as EAV/CR and it enhances the EAV framework by defining classes and their relations in the problem domain (Zhang, 1995). The EAV/CR schema in Table 4.3 is an example where the super-class and their attributes are defined and held in the fields. The Class Hierarchy table shows the relations between the classes. The attribute table stores the records of the class, which the attribute belongs to and their data types. An attribute is the characteristic of a class and it is defined by the attribute type. The instance of a class (aka objects) is recorded in the Object table and the instance fields are recorded in the Data Table, which is similar to the EAV, models (Zhang, 1995).

According to Robert and Fiona, (2007) EAV model has several advantages which, include:

- Flexibility: There are no limits to the number of attributes can be captured per entity therefore the logical database schema can grow without affecting the physical schema.
- Storage: In a clinical database thousands of parameters are available but only a few may be recorded for each patient. The EAV design does not need to reserve space for attributes with NULL values, which may lead to empty (NULL) fields in ER model.

Table 4.3: EAV/CR Database tables an example of the EAV schema with classes and relations (Hellman, 2001).

Class table		
className		
Person		
ClassHierarchy table		
superClassID	subClassID	
Person	Patient	
Attribute table		
ClassID	AttributeName	DataType
Patient	Name	Text
Patient	Date-of-birth	Date

However, the drawback of EAV/CR design is that the system administrator must have excellent understanding of the object-oriented technology in order to design the EVA classes. An EAV/CR database is therefore not user's friendly and flexibility for the average clinician or researcher (Hellman, 2001).

To retrieve information from the EVA/CR tables require complex SQL, which includes the use of INTERSECT operation or self-join for each attribute, where self-join means a join of a table with itself. Besides the complexity of retrieving data from EVA/CR tables and the unfriendly nature of EAV to the end users, the querying operations and performance are slower than the SQL query statements used in conventional ER model (Johnson et al., 1997).

- Limitations of Entity Attribute Value/Class Relations:

The limitations and challenges encountered in the use of Entity Attribute Value/Class Relations (EAV/CR) was researched and fully documented in "Exploring Performance Issues for a Clinical Database Organized Using an Entity-Attribute-Value Representation" by Roland et al., (2000).

Roland et al., (2000) conducted series of experiments using EAV/CR and the conventional ERD database. They concluded that although EAV/CR has a lot of benefits, it has many performance limitations which include complex query structure and poor performance (Rolands and Chen, 2009).

To address the poor query performance of EAV/CR the following were suggested by Robert and Fiona, (2007). Transform the EAV data schema into many conventional tables and make the query design ease for the end users to write simple SQL. Use query optimization techniques to increase the efficiency and breakdown the complex SQL statements into smaller parts to increase the query speed (Johnson et al., 1997) (La et al., 2005).

According to Nadkarni et al. (2001), the speed of retrieving data from conventional databases (ERD) using SQL queries is faster than retrieving data from an EAV/CR. Again, the performance of EAV/CR can be improved by using powerful hardware with extra storage memory. This discussion confirms the research gap that exist in Data Re-representation technique do not address the information overload and optimisation problems in CIS.

4.2 Current Techniques for modelling Clinical Decision Support Systems

The techniques used to model and design decision support systems (DSS) in health information systems are Outranking, Analytical Hierarchy Process (AHP) and Multi-attribute utility theory-MAUT, which were mentioned in Chapter 3. The operations and limitations of the DSS techniques were analysed before the new CMAUT technique and the UML class diagram Data Re-representation are discussed.

- Outranking - Multiple Attributes Decision Making Technique:

Outranking is a decision making technique used in Software Engineering and health applications. According to Fenton et al., (1997), Outranking is a process of solving problems by ranking the solutions from the best to the worst in order to identify the best option. In this technique, outranking relation is defined as a binary relation of a set of actions, where the two actions are (*a* and *b*). According to Roy, (1996), there are two steps in the implementation of Outranking, which are the following:

1. Step1: is to develop and establish the outranking relation between two actions for decision making (Sanderson et al. 2006).
2. Step 2: is the evaluation of the relation between the two actions in the problem domain by using statements (Fenton et al. 1997):

Again, according to Sanderson et al. (2006) the calculation of the concordance index is complex and time consuming. This is because when the numbers of attributes in the problem domain are increased the technique does not function correctly. In this research, the outranking technique is not used because although it reduces the assumptions that are made when handling multi-attribute utility problems, the procedures are implemented differently by different groups hence there is no commonalty (Fenton et al, 1997).

- Analytical Hierarchy Process (AHP)

The AHP technique was developed by Saaty, (1980) to resolve problems that are related to multi-attribute utility (MAUT). AHP begins by representing the problem to be solved as a hierarchy tree, where the top node is the main objective problem while the bottom nodes are the actions to be taken. The decision maker starts from the top node and at each level of the hierarchy makes a pair-wise comparison based on how much contribution each node gives to the next higher node that it is connected to. The pair-wise comparison conducted using either the preference ratios or importance ratios, which are evaluated using numerical scale namely numbering system of 10 or percentages (100%) (Fenton et al., 1997).

Another approach is to form a pair-wise comparison matrix that comprises of criterion in the rows and their attributes in the columns. The values in the matrix are normalised by multiplying each value by the weights recommended by the decision maker. Finally the normalised attributes values are arithmetical summed up and the values added up to 1.0 (Sanderson et al., 2006).

Advantages and Limitations of AHP

AHP was used to conduct the reliability safety assessment of programmable logic components by NASA. Again, NASA uses AHP to determine and choose the best safety feature of the space shuttle (Fenton et al. 1997).

However, the limitations of the AHP method include the following: AHP was criticised because it permits the decision maker to select inconsistent set of weights when normalising the matrix. This denotes that if two criteria are given equal weight then the weighting of the two criteria will be the same in relationship to the others criteria in that same problem domain (Sanderson et al., 2006).

In analytical decision model, the performance ratings are converted into preference scores. In this research the measured preference values are converted into utility unit using a new proposed utility function. The AHP method uses weights that are selected based on the assumptions that they are subjective and are not supported by scientific methods (Sanderson et al., 2006). This research uses statistical approach that is based on logistic binary regression, which is used to generate the beta coefficients from the real life data (HSE, 2006).

4.3 Multi-Attribute Utility Theory-MAUT

According to Fenton et al. (1997), in medical application, all problems are either multiple-objective or multi-attribute. Therefore these problems have to be solved using techniques that consider the multi-dimensional nature of the medical problem. Multi-attribute utility theory (MAUT) was developed to allow the decision maker to put different utility values (i.e. unit) on each of the multi-attribute in the form of weighting. According to Sanderson et al, (2006), this procedure is difficult and not applicable in medical applications because different weights or utility units must be computed for the multiple attributes in the problem domain.

4.3.1 Implementation of Multi-attribute Utility Theory:

In this research, the Multi-attribute utility theory (MAUT) is used for the implementation of the CDSS because it takes into consideration the constraints in the problem domain. The technique also addresses the subjective judgement that decision makers encounter when selecting the best option from different alternative solutions (Sanderson et al. 2006 pp 115). However, in MAUT the decision maker needs to maximise or minimise the objective function that has the various criteria or attributes (Fenton et al. 1997). The procedure for implementing the MAUT technique for decision making is as follows:

1. Define the relevant attributes and their alternatives in the problem domain.
2. Evaluate each attribute and remove irrelevant or dominated alternatives.

3. Assign relative weights to each of the attributes and determine their utility.
4. Sum the attribute weights and determine their overall utility to evaluate each alternative.
5. Perform sensitivity analysis and make a decision.

The procedure for the implementation of MAUT is as follows: first determine the utility function U , which is the sum of all the individual function U_i based on their g_i values. This is then transformed into unit interval $[0,1]$ that is used to form a matrix, in which each column adds up to one (Fenton et al., 1997). One method is to define the function U_i , by allocating weights to each of the attributes in the problem domain and the total value of the weights should sum up to 100%. An alternative approach is to define the function U_i , as the transformation onto the unit matrix $[0,1]$.

The utility of each attribute in the problem domain is determined by using an option where the greatest utility should be ranked top by the decision maker and selected as the optimal solution (Sanderson et al., 2006). The final result is determined by summing the utility functions of each of the attribute under consideration using the following expression:

Total utility function $f(UT) = f(U1) + f(U2) + \dots + f(Ui)$. The total utility function is solved using optimisation techniques.

- Optimisation of utility function in MAUT:

Optimisation is the process of writing a new algorithm or modification of exiting algorithm to improve the efficiency of data retrieval, processing or performance of a system. According to Fenton et al., (1997), in multi-criteria decision making, the key problem is optimising the f function. Optimisation is the process whereby the optimal value of the objective function f is determined in order to satisfy the conditions specified in the problem domain. In the problem domain the objective function f optimises one or more attributes of the elements, A_i , where the A_i , maybe the objects, decisions, candidates etc. In this research, optimisation is defined as the process where one or more attributes of the elements of A_i in the $f(Ui)$ function are optimised in order to satisfy the specified constraints in the problem. Therefore the total utility function $f(UT) = f(U1) + f(U2) + \dots + f(Ui)$ is optimised using Linear Programming Optimisation Technique to determine the optimal solution (Fenton et al. 1997).

4.3.2 Challenges of Multiple Attribute Utility Theory;

The main advantage of MAUT is that it brings order and consistency in solving problems with multiple criteria and multiple attributes (Fenton et al, 1997). This technique permits transparency when allocating weights to each criterion in the problem domain. Another advantage is that MAUT gives the decision maker the ability to select the optimum solutions for a problem with many objectives and multi-criteria. It also allows the user to put different scores or weights on the multiple attributes in the problem domain to determine the best optimum solution. This technique has been used in many applications but not in CDSS (Sanderson et al., 2006).

The disadvantages of MAUT model include the difficulties involved in applying the appropriate weighting techniques because of its subjective nature weight allocation. The utility method is subjective and is defined differently by different people and therefore it is difficult to implement the MAUT in medical application (Sanderson et al, 2006). Sequential elimination is easy to implement, however it has many trade-offs when short-listing the most optimal solution as discussed in chapter 3. This makes it difficult to implement the Sequential elimination and MAUT techniques in clinical decision support systems (Fenton et al. 1997). Other challenges associated with the implementation and applications of MAUT are data representation and construction, which are discussed below:

- Representation of problem: the user must determine the properties or attributes in the problem domain that must meet or satisfy the decision maker's preferences, such that it can be represented by a function, which has a prescribed analytical format (Fenton, 1997).
- Construction of problem: deals with how must the maximisation or minimisation of the objective function be formulated or constructed such that the optimal values can be determined and the various parameters estimated (Fenton et al, 1997).

To address the above challenges, this research has adopted a new system which is known as Combinatorial Multi-Attribute Utility Theory that is discussed in section 4.5 below.

4.4 Proposed new Clinical Data Re-representation using UML and CMAUT

The research gap that is work addresses is the creation of a new clinical data representation mechanism with an optimisation algorithm for the implementation of clinical decision support systems for cardiovascular disease (CVD). The remaining chapters of this Thesis will tackle the issue by proposing the following:-

1. That problem in the clinical domain can be modelled using the class diagram in the UML and formulated it with CMAUT and logical connectors (AND/OR). In this research, this technique would be known as CMAUT Clinical Data Re-representation mechanism.
2. The Construction problem is addressed by developing an algorithm, which has objective utility function that is optimised using LP techniques subject to the constraint identity matrix or unit matrix.
3. The output of the optimisation function can be mapped to the various attributes, which are the nominal standard values and the optimal value is converted into percentage value.

The above procedure will be used to model and build the prediction framework, which consists of a clinical class Re-representation model and CMAUT mathematical formalisation. The prediction framework uses LP prediction algorithm to determine the CVD percentage risk, the comparative output attribute variables and their interpretations.

4.4.1 The New Clinical Data Re-representation Mechanism

In this research, the class diagram is used to model the multiple attributes and the association between the human organs in order to create the new re-representation mechanism (Edoh et al, 2011). The human organ re-representation uses the class diagram as a conceptual model and the CMAUT theory to express problem applying formal specification technique to avoid ambiguity. CMAUT is a decision making tool used in econometrics but it is not utilised in medical application because of its complexity (Sanderson et al. 2006).

In the NL7 CIS, class diagram is used to model and represent the relevant classes required to handle patient-centred activities in the problem domain. The NL7 type of modelling does not focuses on disease as a concept nor captures the organs, their relationship and the multiple attributes. This is a data representation problem that must be address in future NL7 model.

For instance, Balaa, (2008) attempted to capture the kidney and its attributes using class diagram but the model is not comprehensive. This confirms the criticism that the NL7 does not incorporate a methodology for modelling disease, organs and their relationship in RIM (Taylor, 2006).

This research has developed a new framework for optimizing CIS with emphasis on the diseases, patient organs, their attributes and association between the organs. The framework subsumes that a patient is an object, which is made up of six subsystems namely the respiratory, cardiovascular, neurological, coagulation, hepatic and renal as used in the Sepsis-related Organ Failure Assessment (SOFA) algorithms (Vincent et al., 1996).

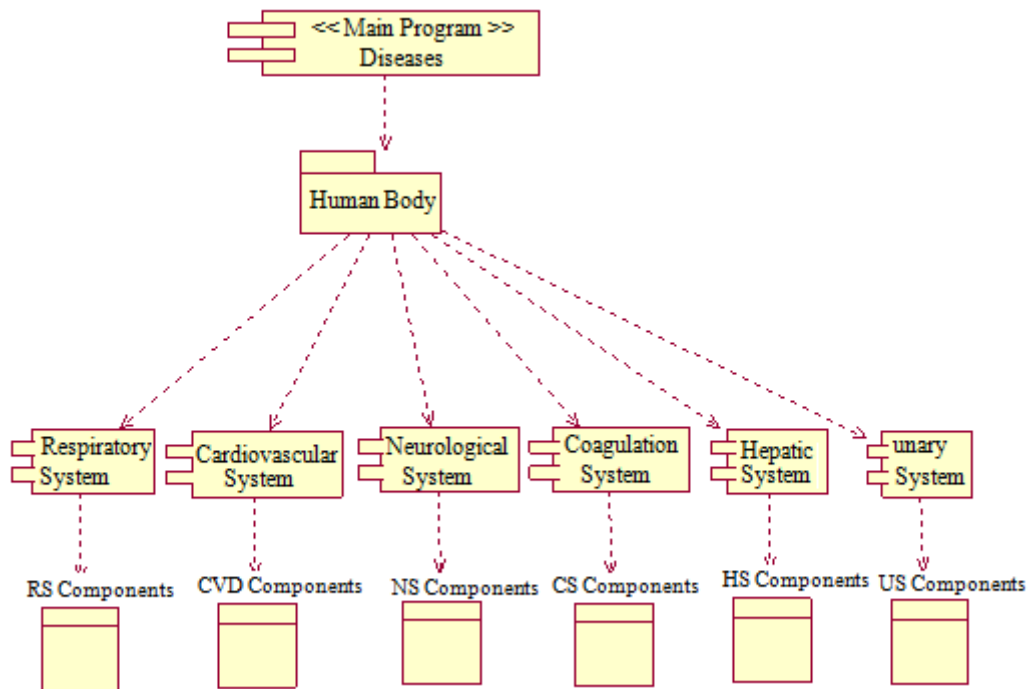


Figure 4.5: Clinical Disease representation using the CMAUT Framework

SOFA is a decision making algorithm used in Intensive Care Unit (ICU) to determine the malfunctions subsystem in patient's organs (Ceriani et al., 2003). The SOFA representation in Figure 4.5 is an extension of the CMAUT concept and UML diagrams. In SOFA the subsystems complement each other in their operation to achieve the requisite homeostatic of the human body. In the CMAUT framework, organs that complement each other in their operation to achieve their predefined goals are called complementary organs. These organs are linked with the AND connectors and their prescribed goals are known as utility function. Likewise organs that can substitute each other in their operation in order to achieve their utility function are known as substitutable organs and are represented with an OR connectors.

In figure 4.5, a disease can affect a patient class, which is made up of six (6) subsystems where each subsystem complements the other and are connected with the AND logical connector. In this conceptual model the association between organs are represented with combinatorial logical operators (OR) and (AND) while the attributes are captured with multiple attributes (MA) theory is called combinatorial multiple attributes (CMA). For instance a hypertension disease that affects the organs; heart, kidney and brain while complementing each other to ascertain that the correct amount of blood and pressure flows through the body to achieve homeostatic are known as complementary organs. Again, pair organs such as kidneys, liver, eyes are substitutable organs because when one is malfunctioning the other act as a substitute to achieve their utility function (Edoh et al., 2011).

The CMAUT conceptual model uses the class diagram in O-O methodology to model the human organs and their relationship and also capture the multiple attributes of the organs. This data representation can be converted in programming language using the formal methods Z or VDM approach (Charatan and Kans, 2004). The CMAUT approach is unique because; 1. It allows CIS to be optimised in order to determine the appropriate amount of data that can be mapped and retrieved for disease assessment. 2. It enables clinical data to be analysed using mathematical based algorithm. 3. It can be used as an epidemiological mechanisms to assess the risk of having CVD disease such as Framingham, QRISK and ASSIGN methodology; 4. It can be expanded to include newly discovered attributes in the framework because of its multiple attribute nature (Ceriani et al., 2003).

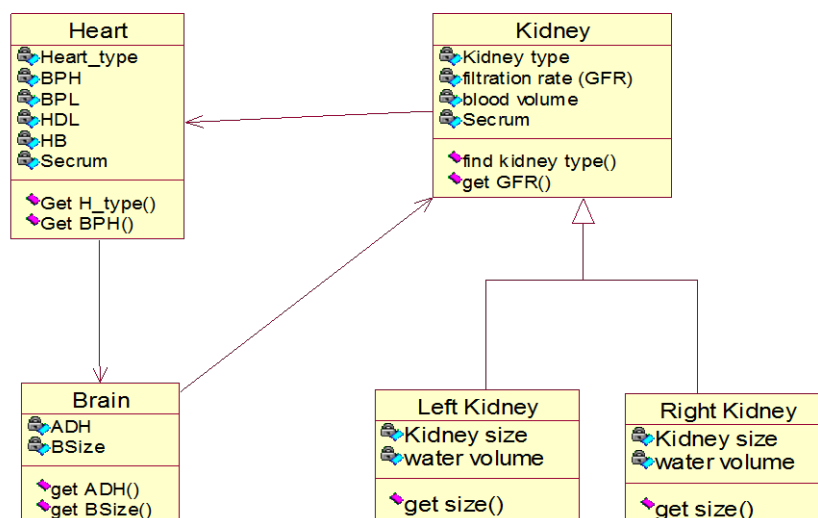


Figure 4.6: New CMAUT clinical Data Re-representation (Imafidon et al., 2009)

4.4.2 New Clinical Data Formalisation using CMAUT for CDSS

The new CMAUT Data Re-representation uses propositional logic and combinatorial human components and their multiple attributes to model the disease. Each component or organ in the human body is classified as an object. This is based on the following assumptions.

Assumption 1: - Clinical human objects, Entities and Diseases

1. In Software Engineering, an object is defined as anything that information can be stored about it and it must have a name, attribute and functions (Fenton et al, 1997). Therefore in this research, every component or organ in the human body is known as an object. However, the human objects have multiple attribute values; hence they are multiple attribute objects. Again, all human objects have names and functions which, includes get heart size, replace knee etc.
2. Entities in this research are defined as two dimensional containers in which data are stored in column (aka field or attributes) and rows (aka records).
3. Disease is defined as an object or the component parts of the body in the problem domain which measured attribute values are different from the specified acceptable standard value.

This new clinical Data Re-representation technique is based on the notion that problems in any domain can be modelled and represented using propositional logics (Wenting, 1995). In propositional logic the main components are represented by literals and they are linked together using logical connectors such as AND, OR, XOR. Examples of these applications are discussed in Wenting, (1995), where problems in e-auction domain are described using propositional logic and the results manipulated before they are transferred into Conjunction Normal Form (CNF) expressions (Edoh, 2004).

The optimal value of a combination of components is determine by using LP techniques after the maximum output CNF expressions have been analysed and converted into set of inequalities (i.e. algebraic calculus). The CNF expressions in the proposed clinical Data Re-representation are transformed into algebraic calculus using the Table recommended (Raman and Grossmann, 1994). In this research constraint matrix in the form of Inequalities Table recommended by Raman and Grossmann, (1994) are used and discussed in chapters 5 and 6.

The first step of the clinical Data Re-representation technique is the application of the combinatorial multiple attribute concept (aka CMA) (Sanderson et al. 2006). In this technique, diseases and their symptoms are described using propositional logic and multiple attributes. The various attributes in the combinatorial expressions are converted into common utility unit using the MAUT (Edoh, 2005). To demonstrate the operation of this new Clinical Data Re-representation technique, disease is defined as the abnormal behaviour of the human organ. Also disease is when the attribute values of the human clinical object under consideration are different from the specified values or norm. The body parts or objects in the problem domain (i.e. disease) are described using propositional logic, which depict the relationship between the various body parts affected.

Assumption 2: - Combinatorial multiple attribute utility theory (CMAUT) for CDSS.

This assumption is based on the fact that every human organ or body part has many attributes, which are associated with their behaviour. In clinical application these attributes are used to define and describe the symptom of a disease (Deutsch et al., 1994). Some of the clinical measurable attributes used in this research are the continuous variables, temperature T , heart rate (HB), blood pressure BP , resistance R and volume V .

In this research, utility function $f(U_i)$ is the service that an organ or combination of organs provide as a contribution to the total output of human body (Deutsch et al. 1994). Therefore the utility function of an organ or combination of organs is determined using the attribute values of each organ (Fenton et al., 1997). Utility unit (U_i) is an instance of the utility function and it is determined by using utility function of each human object or group of objects and the multiple attribute utility theory (MAUT) concepts. Therefore the utility unit (U_i) of each human object is calculated as the ratio of the difference between the measured attribute value of object and the specified acceptable standard value of the object.

The utility unit identifies the relationship between an organ's attribute value from the specified standard value in clinical practice (NICE, 2006). It is subsumed that each of the object's attributes can be converted into common utility function, which is used to denote the behaviour of the organ. For complementary organs, the total utility unit is the arithmetical sum of the utility units of all the attributes of each organ in the disease domain.

However, the total utility unit of substitutable objects in a combinatorial is the sum of the individual object's utility units and not the arithmetical total sum of all the attribute values. This is because substitutable human objects work individually to achieve their utility function.

Assumption 3:- Relationship between clinical human objects and propositional logic

- Complementary organs:

The third assumption is that, all parts of the human body have a relationship with each other, which are complementary, substitutable or high valued. The complementary organs assist each other in performing their duties and are therefore linked to each other using the AND connector. For example, combinatorial components with multi-attributes (CCMA) expression using logic connector AND is written as:

$$X_1: [(C_1 \wedge C_2), P_1, S_1, P_2, S_2] \quad (4.1)$$

In the expression (4.1) the disease X affects the body parts C_1 and C_2 where C_1 and exhibit the attributes P_1 , and S_1 , while organ C_2 has attributes P_2 and S_2 . This research focuses on complementary organs because most of the parts human body complement each other rather than substitute them as discussed in chapter 5 and 7. The generic expression for combinatorial components with multiple attributes (CCMA) using AND connector is:

$$[(C_1 \wedge C_2, \dots, \wedge C_n), P_1, S_1, \dots, P_n, S_n] \quad (4.2)$$

Similarly, the generic expression for two or more body parts which complement each other but has only one attribute is called combinatorial components with single attribute (CCSA) and it is written as

$$[(C_1 \wedge C_2, \dots, \wedge C_n), P_1, P_1, \dots, P_n] \quad (4.3)$$

- Substitutable organs:

Substitutable organs are the body parts or objects that work together and can operate in place of the other or can work as replacement organs for the others particularly when one of the pair of organs is malfunctioning. These organs are called combinatorial substitutable objects or organs and are connected with the OR connector. In the human body, substitutable organs are designed such that when one fails the other can work on their behalf (Guyton and Hall, 2006). Some of the organs are the pair of noses, ear, kidneys, lungs, ovaries, legs, eyes, hands, urethras (i.e. tubes), adrenal gland and gallops tubes lobes among others. Therefore when two body parts or organs act as replacement for each other and have many attributes they are expressed as:

$$X_2 \equiv [(C_2 \vee C_3), P_2, S_2, P_3, S_3] \quad (4.4)$$

When combinatorial substitutable organs have only one attribute, then they are known as combinatorial components with single attribute (CCSA) and it is written as $[(C_1 \vee C_2, \dots, \vee C_n), P_1, P_2, \dots, P_n]$. In this expression, the single attribute P_i is the pressure in each of the objects. The generic expression for combinatorial components with substitutable and multiple attributes (CCMA) is written as

$$[(C_1 \vee C_2, \dots, \vee C_n), P_1, S_1, \dots, P_n, S_n] \quad (4.5)$$

- Highly valued substitutable organs using XOR connector – Brain or the Heart

In CMAUT clinical Data Re-representation technique, the high valued parts such as the heart or the brain are connected to other organs with the “exclusive or” (XOR) connector (Wenting, 1995). Medical research has established that a person is clinically dead when their brain ceases to function however other school of thought argues that it is only when the heart ceases to function that a person is clinically dead (Guyton and Hall, 2006).

This research will not cover this issue but will assume that the two high valued organs are the heart and the brain. Therefore the following example subsumes that the brain is the only high value part of the body while the other organs assist or substitute each other. Therefore for instance a disease X that affects the brain C_1 and the heart C_2 is written as:

$$X \equiv [(C_1 XOR C_2), P_1, S_1, P_2, S_2] \quad (4.6)$$

In this research with the exception of dead brain, the connector XOR will not be used because the patients are considered clinically dead. Thus the focus would be on complementary and substitutable body parts or organs only. CMAUT expression for XOR is written as follows:

$$[(C_1 \text{ XOR } C_2), P_1, S_1, P_2, S_2] \quad (4.7)$$

$[(C_1 \text{ XOR } C_2) \Leftrightarrow U_3]$. The expression (4.7) indicates the common utility value of the two highly valuable organs when their individual attributes i.e. blood pressure and volume are combined it gives a total utility value of U_3 .

4.4.3 Conversion of Multiple Attributes into Utility Unit in CMAUT

The attribute values used in this research are the measureable parameters, which were discussed in chapter 3 and 5. The CVD clinical data is from the HSE, (2006) report, peer reviewed literature and the medical source (NICE, 2006). All the standard clinical values used in this research are from the NICE documents and other literature discussed in section 3.

- Methods of Allocating weight to attributes in CMAUT Combinatorial.

There are two main methods of allocating weights to each attribute in multiple attribute data sets. These are explained in the section below.

Method 1: The weight ranking system (aka performance preference) is a process where the medical experts recommend appropriate percentage or weight that must be allocate to each attribute in the disease domain. This depends on their individual preferences but the arithmetical sum of all the attributes in the combinatorial must add up to 100%. This concept is used in risk factor categories analysed by Wilson et al, (1998) and in Risk factor scoring technique (Hence, 2003). However, Beswick and Brindle, (2006) cautioned that decision made by health experts using scoring and preference allocation values lead to overestimation.

Method 2: this is the use of binary logistic regression approach; this is also referred to as the statistical method of building prediction models (Pencina, 2009). In this method each attribute is allocated β coefficient value that is generated from the regression structural equation obtain by statistically analysing a given data. The study case in this research uses the HSE, (2006) data set and SPSS to analyse and generate the β coefficient values.

According to Terrin, (2003), the use of logistic regression to develop prediction model gives accurate results as compared to the application of Artificial Intelligence (AI) methods.

To convert the multiple attributes in CMAUT expression into a utility unit and at the same time take into consideration the weight each attribute must have in the CDSS of CIS the following equation is proposed.

$$U = \sum [w_i f(s)] \quad \text{where } f(s) = \frac{P_i - P_o}{P_o} \quad (4.8)$$

P_o is the estimated value and P_i is the actual measured and recorded value in the expression.

$$f(s) = P_i - P_o = +^{ve} \text{ normal expression}$$

$f(s) = P_i - P_o = -^{ve}(\text{negative})$ or $P_i < P_o$ the expression will have negative value.

If $f(s) = P_i - P_o = 0$ then $s = 0$; which means this attribute will not be in the objective function because any weight allocated to it will have a resultant zero value.

- Example of converting CMAUT expression into utility unit:

The expression below is for complementary organs with multiple attributes as shown in (4.1).

$$[(C_1 \wedge C_2 \wedge C_3), P_1, V_1, T_1, P_2, V_2, T_2, P_3, V_3, T_3]$$

Using the principle of converting attributes into utility units (4.8) the following are obtained:

$$\begin{aligned} U_1 &= \sum w_p f(sp_1) + w_v f(sv_1) + w_t f(st_1) \\ U_2 &= \sum w_p f(sp_2) + w_v f(sv_2) + w_t f(st_2) \\ U_3 &= \sum w_p f(sp_3) + w_v f(sv_3) + w_t f(st_3) \end{aligned}$$

The w_p is the weight for the pressure, w_v , w_t are the weight for the volume and Total cholesterol. Now the CMAUT expression in the expression framework above can be transformed into mathematical Mixed Integer Program (MIP) for evaluation. In Chapter 6.2, an algorithm is developed that converts the expression into conjunctive normal form (CNF). The CNF expressions are then translated into set of inequalities for evaluation. The expressions are evaluated using LP techniques. In chapter 5 and 7, the simplex LP technique in MATLAB 7.x is used to optimise the objective function generated from the above expressions.

4.5 UML Clinical Data Re-Representation with CMAUT formalisation

To illustrate the CMAUT operations the logical expression for the complementary organs G_1 and G_2 with the attributes as P_1 and P_2 are first written as $(G_1 \wedge G_2, P_1, P_2)$. The substitutable organs, which is G_1 or G_2 with the attribute P_1 and P_2 is expressed as $(G_1 \vee G_2, P_1, P_2)$. Two examples from two different clinical application domains are explained below:

Application of UML and CMAUT to two domain scenarios

The domain scenarios used to illustrate the application of UML class model and CMAUT in CIS are Kidney and Heart related diseases. These scenarios were selected because they clearly depict the relationship between the different organs and their operations. According to Guyton et al. (2006), the two kidneys can work independent and therefore form well-defined actual subclasses with an abstract super-class known as kidney.

Similarly, the main organs that contribute to cardiovascular disease (CVD) are identified as heart, kidneys and the brain. These organs have been modelled in different mathematical formalism in order that they can be simulated for research purposes. However, they have not been modelled using CMAUT and class diagrams. Therefore this research focuses on how the organs that complement each other can be designed and modelled to create homeostatic in the human body during their operation as discussed in Guyton et al., (2006).

4.5.1 UML Clinical Data Re-representation of Kidney Diseases:

Scenario 1: the kidney disease involves the two kidneys working together to perform the function of regulating the flow and extraction of liquid in the human body. As stated in section 4.4.2, each kidney can replace or substitute the other during their operation. According to Guyton et al. (2006), kidney related diseases are acute renal failures and chronic renal failure. This research work focuses on interregional acute renal failure, which is the result of abnormal behaviour of each kidney. In the class diagram Figure 4.7, the kidney is a super class with two subclasses which are the left and right kidneys that are linked with the OR connector. The subclasses inherit the characteristics of the super-class as shown in Figure 4.7.

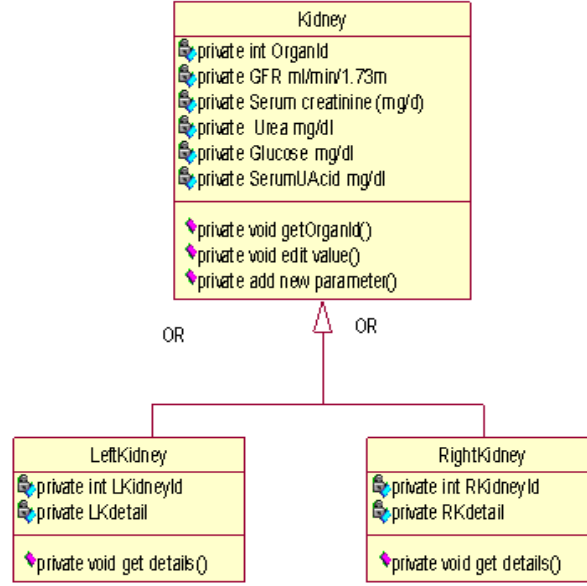


Figure 4.7: Clinical Data Re-representation of Kidney using UML (Edoh et al., 2011)

In Figure 4.7, the two sub-kidneys can supplant each other and have many attributes, hence they are expressed in CMAUT data re-representation using the OR connectors as:

$$X_2 \equiv [(K_1 \vee K_2), P_1, S_1, P_2, S_2] \quad (4.7)$$

In expression (4.7) $K_1 K_2$ are the organs and P_1, S_1, P_2, S_2 are the attributes. The generic expression for these substitutable organs with multiple attributes is written as $[(C_1 \vee C_2, \dots, \vee C_n), P_1, S_1, \dots, P_n, S_n]$.

For substitutable organs, there are two representations

1. At the junction where the two subclasses (left and right kidneys) meet the total utility unit of the kidney is the arithmetical sum of the individual attributes. This is used to form the objective function to be optimised and it is written as

$$\Sigma K = \sum_i^n (U_1 X_1 + U_2 X_2 + \dots + U_n X_n)$$

2. For the distinct individual subclasses the utility unit of the individual attributes are not arithmetical summed together but remain as independent attribute with individual utility unit.

For example there are two separate kidneys; the left kidney has its attributes and utility units while the right kidney has separate attributes and utility units. Therefore after computation the objective function for each kidney will be written as follows:

The objective function for Left Kidney = $\Sigma KL = \sum_i^n (U_1X_1 + U_2X_2 + \dots + U_nX_n)$.

The objective function for Right Kidney = $\Sigma KR = \sum_i^n (U_1X_1 + U_2X_2 + \dots + U_nX_n)$.

The objective function for the Right Kidney and Left Kidney are optimised subject to the unit constraint matrix, using the 1 and 0 mentioned above. The optimisation algorithm is written in MATLAB and it is used to determine the optimal value, which is the percentage risk. Again, the input attributes values can be mapped to the output variable from the framework to identify the attributes in the combinatorial, which attributes need to be analysed and focused on for further medical investigation. This area of study is not covered in this research and it is recommended for further works.

4.5.2 UML Clinical Data Re-representation of Cardiovascular Diseases (CVD):

Scenario 2: this research focuses on hypotension and hypertension diseases, A hypertensive case is described as hypertension (G) is caused by the high rate of pumping of the heart (H), which creates excessive blood pressure on the walls of the arteries (A) and sends appropriate signals to the brain (B) to regulate the flow of fluid in the kidneys (K); Hence the hypertension disease affects three primary organs which are the heart, kidney and the brain component (aka Ant-diuretic hormone ADH), which complements each other in their operations. Figure 4.8 depicts a class diagram with the three main organs and their attributes. The association between them is represented with the AND operator. In this model the combinatorial organs with multi-attributes is expressed using logic connector AND in CMA data re-representation as:

$$X_1: [(H_1 \wedge K_2 \wedge B_3), P_1, S_1, P_2, S_2, P_3, S_3] \quad (4.8)$$

In the expression (4.8), the disease X_1 affects the body parts H_1 , K_2 and B_3 where H_1 exhibits the attributes P_1 , and S_1 , while organ K_2 has attributes P_2 and S_2 etc.

The generic expression for combinatorial clinical organs with multiple attributes using *AND* connector is $[(C_1 \wedge C_2, \dots, \wedge C_n), P_1, S_1, \dots, P_n, S_n]$. The logical expressions (4.7) and (4.8) serve as the input to the optimization framework discussed in Chapter 5.

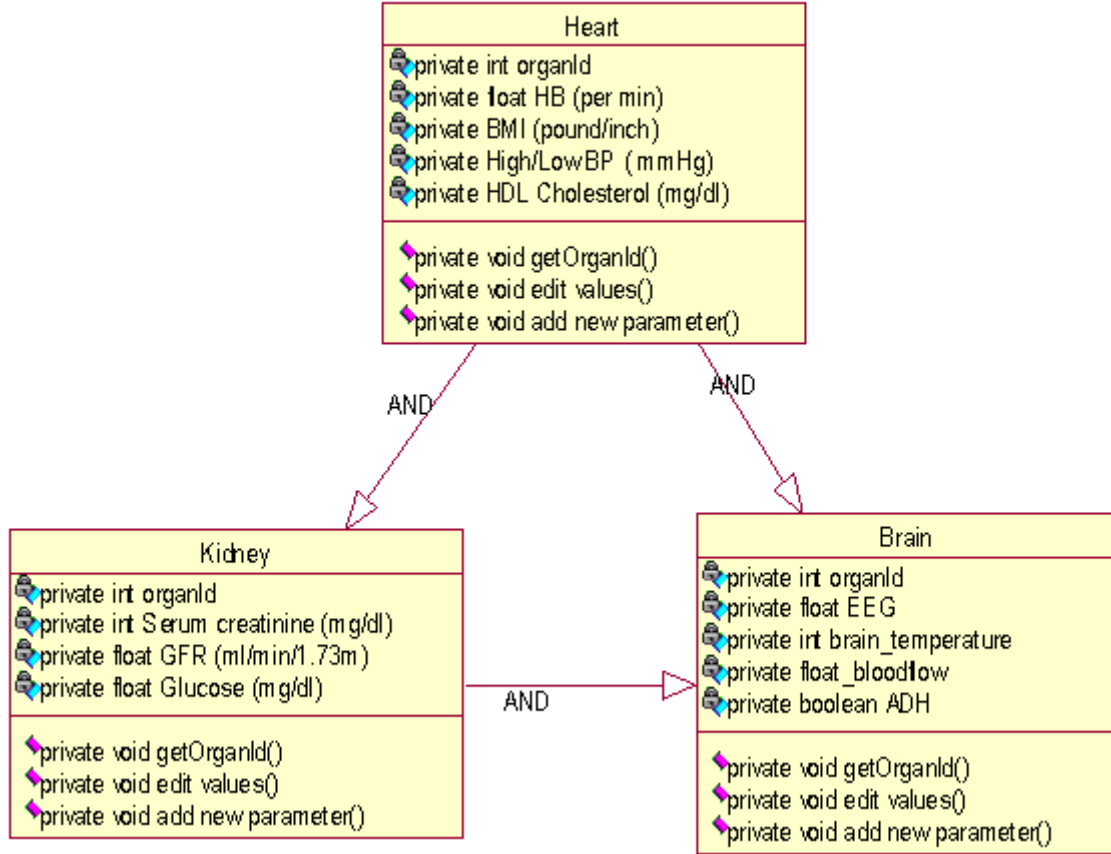


Figure 4.8: Re-representation of heart diseases with UML

Disease X_1 that affects the 3 organs is written as: $X_1: [(G_1 \wedge G_2 \wedge G_3), P_1, D_1, P_2, D_2, P_3, D_3]$.

To convert these attributes into utility function (U) use the procedure as

$$U = \sum[w_i f(s)] \text{ where } f(s) = \frac{P_i - P_o}{P_o} \quad (4.9)$$

The P_o is the standard expected blood pressure while the patient's measured blood pressure is P_i hence the utility unit U_1 of the organ C_1 is shown in expression (4.10) below as:

$$U_1 = \sum(wp1 f(sp1) + ws1 f(sd1) + wq f(sq1)) \quad (4.10)$$

Hence the disease X_1 is expressed in *CMAUT* format as $X_1: [(G_1 \wedge G_2 \wedge G_3), U_1, U_2, U_3]$.

From expression (4.10) it is subsumed that this new algorithm takes into consideration the relationship between expected value P_o of the organ and the actual measured at any time(t).

The baseline value is the standard acceptable value when the organs are performing normally with respect to the person's age, sex, height, weight and they are specified in medical literature (Guyton et al. 2006). The baseline values P_o used for calculating the U_i when the heart is working normally are from Guyton et al. (2006). The systolic and diastolic pressures are as follows; for adults who are over 20 years the values are 140/90mmHg and for diabetics patients the values are 130/60mmHg. According to NICE (2006), in UK the generic standard blood pressure is 140/90 mmHg and the heart rate or heart beat is between 70 to 80 beat per minute. Similarly, the volume of filtrate formed by the two kidneys each minute is glomerular filtration rate (GFR) is about 125ml/min (i.e. 180 litres) a day. The renal blood flow is maintained at constant diastolic pressures of 80 to 200 mmHg. Again since the framework is benchmarked against Framingham algorithm, the ADH in the brain was not measured but specified as Boolean value. Similarly the parameters in the kidney class in Figure 4.8 are not considered. This is further discussed in Chapter 5 and 7 because this research focuses on hypertension, which is a cardiovascular (CVD) disease.

4.6. Summary

This Chapter 4 discusses two medical data re-representation techniques, which are EAV/CR and FOL-ERD. It also reviews the limitations of these data re-representation techniques and how they can be conceptually improved. The EAV/CR model is used for the design of clinical databases. However, it is designed for specific application and it is complicated model with a complex information retrieval mechanism. The chapter also examines decision making models used in CDSS, which include Outranking, Analytical Hierarchy Process (AHP) and Multi-attribute utility theory-MAUT. These decision making models cannot efficiently handle clinical data, which are multi-attribute in nature. These CDSS models are not optimised and therefore cannot solve the information overload problem in CIS.

In this chapter a new Data Re-representation technique that uses UML class model and CMAUT was presented. The application of this new technique to model CVD and Kidney diseases are illustrated and explained in chapter 5. The CMAUT framework can also be used as CDSS model for analysing and predicting hypertension disease.

Chapter 5: CMAUT Optimization Framework for CVD Risk Diagnosis:

5.0 Introduction

The chapter starts with a discussion on the operation of the CVD diagnosis framework using the heart disease scenario described in Chapter 3. The second part of the chapter demonstrates how the framework is used as epidemiological tool to determine the percentage risk of users been hypertensive. As mentioned in chapter 4, this new optimisation framework is made up of two subsystems, which are the Data Re-representation mechanism and the CMAUT optimisation algorithm that uses LP technique to determine the optimal data required for clinical analysis.

5.1 CIS Optimization Framework.

The CIS optimisation framework is made up of the Data Re-representation mechanism, which utilises UML class diagram and the LP optimization algorithm that is designed with the utility function. In this framework, the relationships between the organs in the disease domain are described using logical connectors and the multiple attributes of the organs. This is shown in the class model Figure 5.2 and expressed in CMAUT mathematical format. The weight allocated to each attribute is determined using the beta coefficient values from the SPSS binary logistic regression analysis and incorporated into the expression $U = \sum w_i f s_i$. The CMAUT expressions formulated from the multiple attributes in the class diagram are converted into mathematical format which serves as input to the LP optimisation algorithm.

The procedure used to convert the CMA expressions from the class model into CMAUT mathematical format is discussed in section 4.4.3: The Data Re-representation mechanism and algorithm in the framework works as follows:

The stepwise procedure is:

1. Write the relations to be optimized in the disease domain using logic expression;
2. Group the attributes and calculate the utility function using $U = \sum [w_i f (S_i)]$

3. Use symbolic manipulation, to convert the logical expressions into conjunctive normal form (CNF) using the U for each attribute;
4. Translate the logic expressions in CNF into linear mixed integer variables and set of inequalities (aka constraints) using unit matrix or Raman's Transformation table.
5. Establish the objective function to be maximize or minimize
6. Use the LP algorithm in the framework to optimise the objective function: $f(Ui)$
7. Convert the evaluated value after the optimisation process to percentage
8. Map the optimal X_i values from the optimisation process with the attributes.

The activity diagram 5.1 below depicts the operation of the CMAUT Framework:

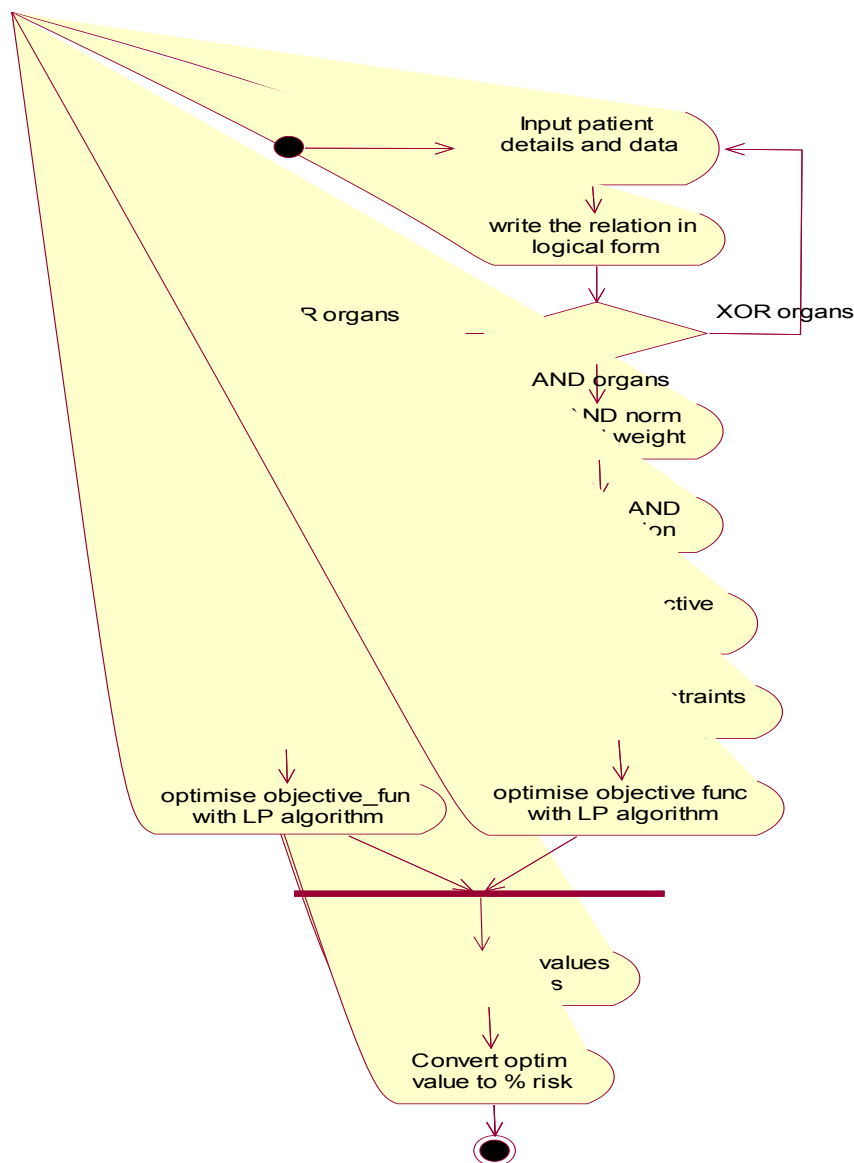


Figure 5.1: Flowchart and activity diagram of the CMAUT Framework

5.2 Data Re-Representation Mechanism using UML

The Data Re-representation mechanism is made up of two components namely:

1. A subsystem for modelling clinical data using UML class diagram and
2. A CMAUT subsystem for the Re-representation of the clinical data in logical format and formalising them for mathematical manipulation and optimisation using LP method.

The first subsystem uses class diagram to model the organs, their multiple attributes and the relationship between the organs in the disease domain. In class diagram, the association between the complementary organs are modelled using the AND logical connectors. Substitutable organs on the other hand are shown as subclasses that inherit the attributes of the super-class, which is an abstract class in human physiology. The subclasses are the two real kidneys, ears, eyes, noses (Guyton and Hall, 2006). This concept is used in UML class modelling, where the super-class is known as the abstract class because it does not exist. The substitutable organs are represented using the OR logical connector, which shows that the organs can replace each other in their operations (Edoh et al., 2011).

- CMAUT data representation system and formalism

To illustrate the operation of the CMAUT data representation, three complementary organs namely G_1 and G_2 and G_3 with multiple attributes P_1, D_1 , for organ G_1 , P_2, D_2 , for organ G_2 and P_3, D_3 for organ G_3 are used. First the clinical data representation is written in the CMAUT logical expression as the disease $X1$ that affects three complementary organs is $X1: ((G_1 \wedge G_2 \wedge G_3), P_1, D_1, P_2, D_2, P_3, D_3)$. These attributes are converted into utility units using the utility function $f(Ui)$ and the procedure used is as follows:

$$Ui = \sum [w_i f(s)] \quad \text{where } f(s) = \frac{P_i - P_o}{P_o} \quad (5.1)$$

In the expression (5.1), the P_o is the expected pressure and the patient's measured blood pressure is P_i therefore the utility unit U_1 of the organ G_1 is shown in expression (5.2) as:

$$U_1 = \sum (wp_1 f(sp_1) + ws_1 f(sd_1) + wq f(sq_1)) \quad (5.2)$$

This research focuses on hypertension, which is a cardiovascular (CVD) disease. The hypertension disease $X1$ is written in CMAUT format as $X1: [(G_1 \wedge G_2 \wedge G_3), U_1, U_2, U_3]$. Where the U_i values are the utility units of the measurable CVD attributes and risk factors. From expression (5.1) it is evident that the CMAUT algorithm takes into consideration the difference between expected standard attribute value P_o of the organ and the actual measured P_i recorded at the time of clinical examination (t). As stated in section 4. 5, the expected standard value P_o is the measure of the normal performance of an organ with respect to the person's age, sex, height, weight as indicated in NICE, (2006) report and medical literature (Guyton et al., 2006).

Therefore the CVD baseline values P_o used to calculate the utility unit U_i in this research are: For systolic and diastolic pressures of adults over 20years the values are 140/90mmHg and for diabetic patients the values are 130/60mmHg. The acceptable heart rate is between 70 and 80 beat per minute but in this research, the standard value used is 50 beats/minutes. The ADH in the brain was not used because it is not applicable in CVD prediction models. This research focuses on hypertension as a cardiovascular disease (CVD) therefore the demographic and clinical data used are shown in Table 5.2 below.

- Clinical data CVD Re-representation with UML and CMAUT;

The scenario below shows how blood flows through the cardiovascular system (CVD) of the human body (Guyton et al. 2006). This research applies the heart disease scenario to describe the behaviour pattern of hypertension and presents the disease pattern using the CMAUT logic format:

Hypertension disease is described as follows: - Hypertension (G) is caused by the high rate of pumping the heart (H), which creates excessive blood flow with pressure (P) on the walls of the arteries (A) that sends signals to the AHD in brain (B) and the kidneys (K) to regulate the blood flow and maintain haemostatics. The UML class model re-representation of the CVD is shown in Figure 5.2 below.

In this research, two sets of CMAUT models are designed and built using the MATLAB software for diagnosis and prognosis of CVD. The models are built based on the above hypertension CVD scenario and the models are simulated using the HSE, (2006) clinical data discussed in chapter 3.

5.3 Application Cardiovascular Disease (CVD) in the CMAUT Framework

To illustrate the operation of the CMAUT framework, the Cardiovascular Disease scenario discussed in Chapter 3 was re-formulated as follows: that the hypertension disease (G_1) “is caused by” high rate of pumping blood by the Heart (H) that “sends” excessive high pressure blood to the Atrial(A) which “send signal to” the Ant Diuretic Hormone (ADH) in the Brian (B) to regulate the flow of fluid to the kidneys (K). The organs in this scenario are complementary organs because they assist each other in performing their duties.

In this scenario the CVD clinical data of two participants (a female and a male) selected from HSE, (2006) data in appendix 3.4 and shown in Table 5.1 are used:

Table 5.1- The three participants used for illustration and the simulation exercises are:

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
42,831,101.00	No	No	35	Men	White	64.50	18.15	135.00	70.50	1.70	92.00	No	4.10	Yes	No	No

Simulation 1: the data of the two selected participants (a female and a male) from HSE, (2006), which are the Age, Sex, BMI, HB, BPH, HDL, TC and MAP were input into the framework. The framework computes the value of the Absolute Percentage Risk (APR) of each of the participants to verify if they have hypertension and the results are recorded in Table 5.9A and Table 5.9B. Again, the attributes in the combinatorial that have the optimal utility values, are determined and the output variables mapped to the respective attributes as indicated in Table 5.9A and Table 5.9B.

The CVD scenario discussed in chapter 3 is modelled using the UML class diagram and CMAUT re-representation technique in Figure 5.2. The Figure 5.2 depicts the CVD risk factors and multiple attributes used by existing Web CVD risk calculators that are designed with the Framingham equations (Sheridan et al., 2003):

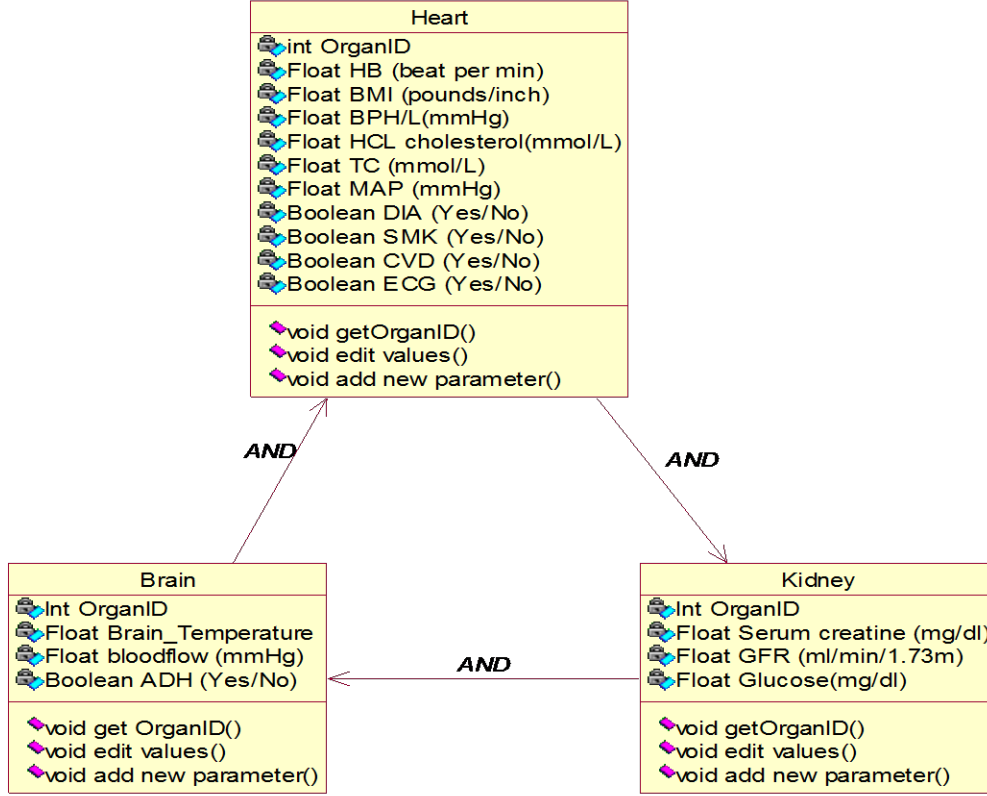


Figure 5.2: Re-representation of CVD risk factors for CMAUT framework

The combinatorial organs and multi-attributes in Figure 5.2 are written in *CMAUT* format using the logic connector AND as follows:

$$G_1: [(H_1 \wedge K_1 \wedge B_1 \wedge A_1), P_{1H}, V_{1H}, P_{2H}, V_{2H}, P_{3K}, V_{3K}, P_{4H}, V_{4H}, P_{5A}, V_{5A}, P_{6A}, V_{6A}, P_{7H}, V_{7H}] \quad (5.3)$$

$$G_1: [(H_1 \wedge K_1 \wedge B_1 \wedge A_1), U_{1H}, U_{2H}, U_{3K}, U_{4H}, U_{5A}, U_{6A}, U_{7H}] \quad (5.4)$$

In expression (5.3), the hypertension disease G_1 affects the organs H_1, K_1, B_1 , and A_1 , which have the attributes P_i , and V_i , that are converted into utility units U_i as in expression (5.4).

The proposed CMAUT framework in section 5.1 was converted into an epidemiological CVD risk prediction model. The new CMAUT framework transformed into risk prediction model by applying the same CVD risk factors as used in existing Web risk calculators and Framingham equations (Sheridan et al., 2003). This approach enables the risk results from the CMAUT framework to be compared with risk results from existing CVD risk prediction tools such as Web calculator and Framingham equations (Chuang et al., 2007). The multiple attributes used in the Framingham prediction model and for the simulation exercises in this research are: HB - R_1 , BPL - P_L , BPH - P_H , MAP - M , BMI - V , HDL - D and TC- T .

As discussed in Chapter 4, for the CVD scenario the brain acts as a catalyst and therefore its attributes are not considered in the CMAUT framework class diagram. Likewise the Atrial system that carries the blood, is not be modelled in the class model. However, the MAP values are calculated as the difference between BPH and BPL and modelled as an attribute of the heart (Guyton et al, 2006). Therefore the expression (5.3) is modified accordingly to (5.5)

$$G_1: [(H_1 \wedge K_1 \wedge A_1), P_{1H}, V_{1H}, P_{2H}, V_{2H}, P_{3K}, V_{3K}, P_{4H}, V_{4H}, P_{5A}, V_{5A}, P_{6A}, V_{6A}, P_{7H}, V_{7H}] \quad (5.5)$$

For the purpose of comparison with other CVD risk calculators, the expression (5.5) is rewritten using Utility Units as shown in (5.6) to reflect only the classes in the UML class diagram and the attributes used for the development of the CVD - CMAUT framework:

$$G_1: [(H_1 \wedge K_1 \wedge A_1), U_{1H}, U_{2H}, U_{3K}, U_{4H}, U_{5A}, U_{6A}, U_{7H}] \quad (5.6)$$

The expression (5.6) is used as the input for modelling the CMAUT optimisation algorithm in the framework and also to determine the percentage risk of a user having hypertension.

5.3.1 The CMAUT Optimisation Algorithm:

The CMAUT logical expression written from the data re-representation mechanism is converted into mathematical formalisation using algebraic calculus that serves as input into the algorithm. For complementary organs, the utility units of the individual attributes are arithmetically sum together to form the objective function. The function is written as

$$Z = \sum_i^n (U_1X_1 + U_2X_2 + \dots + U_nX_n)$$

This objective function is optimised subject to the constraint matrix, generated in the unit matrix formation or Table 5.3. The Table 5.3 depicts the attributes that are measured and used in the optimisation process as 1 and those not used in the inequalities as 0.

5.3.2 Determination of Initial Clinical Absolute Percentage Risk (APR) in CVD

In CVD health care, diagnosis is the first stage of disease management. Therefore in clinical settings, the determination of CVD risk assists in the timely intervention of the cause of the CVD disease, thus the prediction of Initial clinical Absolute Percentage Risk (APR) is essential (Guyton et al. 2006). In this research, Absolute Percentage Risk (APR) is defined as the percentage probability of a participant being hypertensive in the current state based on their clinical measurable data. This absolute percentage risk (APR) is used for early diagnosis of CVD and it is not directly related to time (Panagiotakos and Stavrinos, 2006). In prognosis, predictive percentage risk (PPR) is defined as the measure of the likelihood of a participant developing a disease over a specified time period. Relative Risk (RR) on the other hand is the measure of the chances of risk occurring in two different groups of people (Anderson et al., 1991).

- CVD data used for modelling the CMAUT Framework

From the HSE (2006) report, the demographic and clinical data of all participants who provided detail complete data were filtered out and used in this research. The demographic data used are the participant's serial number (PIND), age, sex, ethinda (i.e. ethnic origin). For benchmarking, the APR results from the CMAUT framework are compared with the results from existing Framingham equation and web calculators. The clinical data used for the development and evaluation of the CMAUT model are HB, HDL, BMI, TC, HDL, BPH, BPL, ECG/LVC and CVD, Diabetic and Smoking. These are shown in the Table 5.1 and Table 5.2.

Table 5.2: Demography and Clinical data used in this research:

Attributes Used in work	Medical name	READ code	Baseline value	Min value in HSE	Max value in HSE
OmpulvalHB	Heart Pulse Rate (Beats Per Minute)	NA	50BPM	20.50	136.50
BMI	Body mass index	22K..%	25.5 kg/m2	13.20	49.66
OmsysvalBPH	Sitting systolic blood pressure	246R.	140 mmHg	84.00	225.00
OmdiaBPL	Sitting diastolic blood pressure	246Q.	90 mmHg	39.00	117.50
Hdlval1HDL	High-density lipoprotein (HDL) cholesterol	NA	1.2 mmol/L	0.50	4.40
OmmapvalMAP	Mean Atrial Pressure	NA	100 mmHg	58.50	140.00
CholvalTotalCholesterol	Total cholesterol measurement	44PH.	5.5 mmol/L	2.10	11.40
Diabetic	Doctor diagnosed diabetes	C10F.%	N/A	NO	YES
Smoking	Current smoker	137R	N/A	NO	YES
ECG/LVC	Had electrical recording ECG of the heart.	NA	N/A	NO	YES
CVD	Family history	12...%	N/A	NO	YES
Bp1	Doctor diagnosed high blood pressure hypertension	G2 and 105	N/A	NO	YES
P_serial No	Participant serial number	NA	N/A	10,102,102.00	82,056,101,00
SEX	Participant sex	-	N/A	N/A	N/A
Age	Participant Age	-	Years	16 Years	92 Years
ethinda	Participant ethnic origin	9i0..% to 916E.	N/A	N/A	N/A

5.4 Modelling of the CMAUT Framework

To transform the optimisation framework to CVD risk prediction model, two types of CVD risk prediction models namely; the Web based heart risk calculators and Framingham Risk equations were discussed in chapters 3 and 8 (Wilson P, 1998) (Brindle 2003). It was identified that these prediction models use both measureable and non-measureable attributes. The CMAUT framework uses only measureable attribute values for the computation of APR. Therefore $X_1, X_2, X_3, X_4, X_5, X_6$ and X_7 , which are measurable attributes that leads to hypertension disease (G_1) were selected see Table 5.2 and Figure 5.3. Table 5.3 is a unit matrix where 1 represents attribute values measured and 0 indicates the attribute measured but not included in the inequality statement. The objective function to be optimized is;

$$Z = \sum_i^n (U_1 X_1 + U_2 X_2 + \dots + U_n X_n) \quad (5.7)$$

Table 5.3: Shows attributes values for organs

Utility Attrib	X1	X2	X3	X4	X5	X6	Xn
HB	1	0	0	0	0	0	0
BMI	0	1	0	0	0	0	0
BPL	0	0	1	0	0	0	0
BPH	0	0	0	1	0	0	0
HDL	0	0	0	0	1	0	0
MAP	0	0	0	0	0	1	0
TC	0	0	0	0	0	0	1

To verify the framework and predict the percentage CVD risk of each participant, the demographic and clinical data from HSE, (2006) were used.

Expressions (5.1) and (5.2) were used to calculate the utility unit for each attribute. First the weight allocated to each attribute was assessed. Two methods are recommended for the weight assessment: For diagnoses, medics may recommend the percentage weight that must be allocated to each attribute depending on their ranking in the disease domain as discussed in chapter 4, section 4.2. Alternatively, the weights are determined with the aid of binary logistic regression where the regression equation $y = a + \beta_1 X_1 + \dots + \beta_n X_n$ is used to allocate the weights for diagnosis and prognosis. In this research, SPSS was used to analyse the clinical records of 3645 participants from the HSE, (2006). The beta coefficient values obtained from conducting binary logistic regression are as follows: a constant value of -10.26, (HBI) 0.211, (BMI) 0.077, (BPH) -0.285, (HDL) 0.200, (MAP) 0.335 and (TC) 0.0766. The beta coefficient values were made the weights of each attribute in the equation.

The beta coefficient values were incorporated into the algorithm developed in MATLAB 7.x that computes the percentage CVD risk of each participant. Figure 5.3 is an example of the output screen of a participant from the HSE, (2006) report. The screen depicts the measured attribute values of the participant which was converted into utility unit using the formula (5.1) and the standard parameters in Table 5.2. The aim of the optimisation algorithm in the framework is to find the attribute(s) in the combinatorial organs that has an overall utility unit that maximizes the utility value to be retrieved for primary healthcare investigation.

The above optimisation problem is presented in a LP format with an objective function as follows: $Max: U_1X_1 + U_2X_2 + U_3X_3, \dots, +U_nX_n$ and optimised subject to the constraint matrix in Table 5.3.

The output of the MATLAB program of the CMAUT optimisation algorithm is shown in Figure 5.3 below. In the program, the standard values of the attributes namely HB = 50, BMI= 25, BPH=140, BPL= 90, HDL= 1.2, MAP =100 and TC = 5.00, which were discussed in chapter 3 and shown in Table 5. 2 in this chapter were used. The output from the CMAUT optimisation framework for the diagnosis of CVD is shown in Figure 5.3 below.

Model_GUI

Percentage Risk for Heart Diseases

Patient serial no. Patient group no.

Patient age Diabetic (y/n)

Gender (m/f) Smoke (y/n)

HB

BMI

BPH

BPL

HDL

MAP

Total Choles

Risk analysis **14.5909**

	X-names	X-values
1	x1	51
2	x2	26
3	x3	1
4	x4	1
5	x5	51
6	x6	6

Figure 5.3: Output screen of CIS_CMAUT framework from desktop.

The LP optimisation algorithm in MATLAB determines the optimal valuation attribute and the maximum value. The results in Figure 5.3 after optimisation are optimal integer values $X_1 = 51$, $X_2 = 26$, $X_3 = 1$, $X_4 = 1$, $X_5 = 51$ and $X_6 = 6$. The solution indicates that the maximum value is 14.59 and the optimal values of $X_1 = 50$, $X_2 = 26$, and $X_5 = 51$ are the attributes that require investigation. Hence through mapping the data that requires investigation are X_1 (HB), X_2 (BMI), X_5 (MAP) and risk factor value is 14.59 %.

The attributes identified for detail investigation are correct if they are compared the attributes values in Figure 5.3 based on the normal values of HB1 = 50.0; BMI1 = 25.5; BPH1 = 140.0; BPL1 = 90.0; HDL1 = 1.20; MAP1 = 100; TC1 = 5.0 applied in the optimisation algorithm.

5.5 Implementation of CMAUT Optimisation Diagnosis Framework models 1 and 2

The CMAUT optimisation model1 and 2 were designed using the following procedure: - first, only the measurable risk prediction factors were used instead of entire non-measurable CVD risk factors discussed in Chapter 3 and 7. This is because diagnosis prediction models are designed to predict the clinical percentage risk probability of a user been hypertensive based on their current measurable clinical CVD attributes (Panagiotakos and Stavrinos, 2006). Therefore the diagnosis prediction model must have the ability to use current appropriate risk factors in the user's records to analysis and predict the absolute percentage risk (APR). Secondly, the binary logistic regression approach was used to model the new prediction diagnosis algorithms because it gives relatively accurate results. The procedure that was applied to create the new CMAUT optimisation models uses the CVD risk prediction factors and SPSS binary logistic regression, which are explained in the sections 5.5.1.

5.5.1 Determination of APR Risk using CMAUT CVD Framework Model 1

- Determination of the APR using the CVD Model1 CMAUT framework

In CVD health care, diagnosis is the first stage in disease management while CVD risk prediction assist in the timely intervention of the cause of diseases hence clinical risk prediction is essential. In this research, the absolute percentage risk (APR) is defined as the percentage probability of a participant being hypertensive in their current state depending on their measurable clinical data. The two diagnosis models built in this research used the beta values in the Variables in the Equation from SPSS logistic regression.

- Model 1 of the CMAUT framework

The design of the CMAUT Diagnosis framework Model 1 was carried out using clinical data from HSE, (2006). The data used are from the 4316 participants, who are over 16 years old.

The CVD data of the 4316 adults were analysed using SPSS to generate the Table that contains the “Variables in the Equation”. This is because the core sample used for the CVD survey was participants who are over 16 years (Craig et al, 2006a).

The procedure for conducting the logistic regression in SPSS is as follows: Open the Excel spread sheet in SPSS then Select Analyse in the main menu. In the drop down menu, select Regression, then click binary logistic regression. When the dialog box appears, select the Bp1 which is dichotomous value because they are made up of either YES or NO as dependant variable. Then select all the other measurable attributes as the Independent values as indicated in Figure 5.4 below.

Point-and-click
 Step 1: Analyse → Regression → Binary Logistic
 Step 2: Select bp1 as Dependent variable and move it to the text box
 Step 3: Select: OmpulvalHB, BMI, OmdiavalBPL, OmsysvalBPH, Hd1val1HDL, OmmapvalMAP, CholvalTotalCholestrol and move them to Covariate text box on the right
 Step 4: Click on Options,
 Step 5: Select classification plot, Hosmer and Lemeshow Test goodness fit, correlation estimate, CI of 95%
 Step 6: Click on OK

Figure 5.4: Procedure for determining Logistic Regression in SPSS

The output from the SPSS shows the variables in the equation but excludes the BPL see Table 5.4 below. The binary logistic regression analysis gives the results in Figure 5.5: It must be noted that OmsysvalBPL is not in the equation Figure 5.5 from SPSS below.

Table 5.4: Classification Table 1(a,b) total 4316 participants

Observed			Predicted	
			bp1	
			0	1
Step 0	bp1	0	3584	0
		1	732	0
Overall Percentage				
				100.0
				.0
				83.0

- a: Constant is included in the model.
 b: The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step	OmpulvalHB	.211	.149	2.018	1	.155	1.235
1(a)	BMI	.077	.010	65.667	1	.000	1.080
	OmsysvalBPH	-.285	.223	1.638	1	.201	.752
	Hdlval1HDL	.200	.120	2.793	1	.095	1.221
	OmmapvalMAP	.335	.223	2.263	1	.132	1.398
	CholvalTotalCholestrol	.076	.041	3.440	1	.064	1.079
	Constant	-10.261	.480	457.216	1	.000	.000

a Variable(s) entered on step 1: OmpulvalHB, BMI, OmsysvalBPH, Hdlval1HDL, OmmapvalMAP, CholvalTotalCholestrol.

Figure 5.5: Output from the logistic Regression in SPSS

- Variables in the regression equation and Weight allocation:

In this model 1, the regression beta coefficients obtained from the SPSS analysis were used as the weights for each attribute. This is applied in the utility formula (equation 5.1) to calculate the utility unit (U) of each specific attribute and the overall utility unit is the sum of all the individual utility units. In the CMAUT model, the objective function which is made up of utility units of each attribute are maximised subjected to the inequalities constraint matrix.

First the attributes values, which are β_i values in Figure 5.5 are converted into utility units using the equation $U = wt (P_0 - P_1)/P_0$. Secondly, in this research the CMAUT model uses the filtered 4316 clinical data records from (HSE, 2006) and with the individual weights from the SPSS binary logistic regression the equation becomes $y = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$. Thus the coefficient of regression (β_i) of each attributes in the equation for over 16years participants are shown in Figure 5.5 above as OmpulvalHB = 0.211, BMI = 0.077 OmsysvalBPH = -0.285, Hdlval1HDL = 0.200, OmmapvalMAP = 0.335, CholvalTotalCholestrol = 0.076 and the constant value = -10.261. The number of attributes in the equation is reduced from seven (7) to six (6) because the OmdymaBPL did not appear in SPSS output table below. Therefore the above the equation can be written as follows:-

$$y = -10.261 + 0.211xHB + 0.077 xBMI + -0.285xBPH + 0.200xHDL + 0.335xMAP + 0.076 xTC.$$

The number of measurable attributes entered into SPSS for binary regression analysis was seven, therefore the equation should be:

$$y = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7.$$

From the assumption made in section 5.4 and explained in equation 5.7, the seven attributes in the CMAUT expression which were used as input into the SPSS for analysis, are:

$$G_1: [(H_1 \wedge K_2 \wedge B_3), R_1, B_2, P_{H1}, P_{L1}, H_1, M_2, T_2] \quad (5.8)$$

However, owing to the reduction of the number of attributes from seven to six the final CMAUT logical expression is rewritten as follows:

$$G_1: [(H_1 \wedge K_2 \wedge B_3), R_1, B_2, P_{H1}, H_1, M_1, T_2] \quad (5.9)$$

In expression (5.8) the subscripts indicate the organ that each attribute is related to, however in CMAUT complementary expression all the attributes complement each other therefore the subscripts can be removed and the utility unit of each attribute calculated as below.

- Conversion of the attributes into utility unit

Example 1:- The participant 2 from the HSE, (2006), has the information in the table below: For computation, the attributes are first converted into utility units by converting each of the coefficients of regression (β_i) into weights using the equation $U = (\beta_i \frac{P_i - P_o}{P_o})$.

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
42,831,101.00	No	No	35	Men	White	64.50	18.15	135.00	70.50	1.70	92.00	No	4.10	Yes	No	No

Therefore the utility unit for each attribute is calculated as follows:-

$$\begin{aligned} U_{HB} &= U_R = ((\beta_i (R_1 - R_0)/R_0)) = 0.211 ((50-34)/50) = 6.75 \\ U_{BMI} &= U_V = ((\beta_i (V_1 - V_0)/V_0)) = 0.077 ((25.5-13.2)/25.5) = 3.714 \\ U_{BPH} &= U_P = ((\beta_i (P_{H1} - P_{H0})/P_0)) = -0.285 ((140-122.5)/140) = -3.56 \\ U_{HDL} &= U_D = ((\beta_i (D_1 - D_0)/D_0)) = 0.200 ((1.2-1.8)/1.2) = -10 \\ U_{MAP} &= U_M = ((\beta_i (M_1 - M_0)/M_0)) = 0.335 ((100 - 100)/100) = 0 \\ U_{TC} &= U_T = ((\beta_i (T_1 - T_0)/T_0)) = 0.076 ((5-5.2)/5) = -0.304 \end{aligned}$$

Box 5.1: Computation of the Utility Units

From the utility unit calculation, the expression (5.9) can be rewritten as

$$G_1: [(H_1 \wedge K_2 \wedge B_3), U_R, U_V, U_{PH}, U_D, U_M, U_T,] \quad (5.10)$$

This expression shows the relative effect and influence each attribute has on the disease in the problem domain. This is represented in utility units in the expression. The calculated U_i vector values in expression (5.10) only indicate the relative strength of each attribute in the disease domain. Hence, to make the expression (5.10) meaningful, it must be converted into a format that can be manipulated and analysis for decision making (Deutsch et al, 1994).

For implementation purpose, the calculated utility unit U_i , of each vector value is attached to their respective variables as $UnXn$ to make the operation meaningful. This is because the organs in the disease domain complement each other and therefore the overall or total utility unit is the arithmetical sum of the individual utility units. Hence the expression (5.10) and the calculated individual utility units with their respective attributes are rewritten as $Z = \sum_i^n (U_1X_1 + U_2X_2 + \dots + U_nX_n)$. This expression is then formulated as the objective function that must be maximized and the constant value (a) in the SPSS regression equation is incorporated into the objective function. The generic objective function is expressed in equation (5.11) as follows:

$$Z = \sum_i^n (a + U_RX_R + U_VX_V + U_{PH}X_{PH} + U_DX_D + U_MX_M + U_TX_T) \quad (5.11)$$

When the individual calculated utility units are substituted into the objective function (5.11), the final function will be as:

$$Z = \sum_i^n ((-10.261) + 6.75X_R + 3.714X_V + (-3.56)X_{PH} + (-10)X_D + (0)X_M + 0.304X_T) \quad (5.12)$$

The objective function (5.12) is maximised subject to the set of inequalities or constraints discussed in section 5.4. In these constraints, each inequality means the measured attribute must be less than or equal to the normal value in order to meet CVD standard set by NICE in their guidelines. For example, the variable X_R for Heart Beat must be less or equal to 50 beat per minute to satisfy the NICE, (2006) condition.

$$\begin{array}{rcl}
\text{[(CX)]_R} + 0 + 0 + 0 + 0 + 0 & \leq & 50 \\
0 + \text{CX_B} + 0 + 0 + 0 + 0 & \leq & 25.5 \\
0 + 0 + \text{CX_PH} + 0 + 0 + 0 & \leq & 140 \\
0 + 0 + 0 + \text{CX_H} + 0 + 0 & \geq & 1.2 \\
0 + 0 + 0 + 0 + \text{[(CX)]_M} + 0 & \leq & 100 \\
0 + 0 + 0 + 0 + 0 + \text{[(CX)]_T} & \leq & 5
\end{array}$$

Box 5.2: Set of inequalities for the constraints matrix

To solve the above optimisation problem the linear programming method in LINPROG MATLAB is used. First all the coefficients in the objective function are formatted as in Box 5.3 below:

$$F = 10.261 + 6.75 + 3.714 - 3.56 - 10 + 0 + -0.304$$

A = the coefficient in the unit constraint matrix;

$$b = [50, 25.5, 140, -1.2, 100, 5];$$

$$1b = \text{zeros}(6, 1);$$

Secondly, the program calls a linear programming routine in MATLAB library.

$$[X, fval, \text{exitflag}, \text{output}, \text{lamba}] = \text{Linprog}(f, A, b, [], [], 1b);$$

Box 5.3: The CVD Optimisation algorithm using linear programming in MATLAB

In Box 5.3, the set of inequalities **A**, in the linear programming function is expressed using identity matrix or unit matrix format as shown below:

$$A = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

Box 5.4: Set of Inequalities unit matrix A used for Optimisation in MATLAB

The unit matrix in Box 5.2 and Box 5.4 were used to develop the MATLAB CVD Optimisation algorithm shown in Box 5.3. The full Optimisation algorithm was written using MATLAB source code shown in Figure 5.6 below. The results of the CVD risk after executing the MATLAB Optimisation program is shown in Figure 5.7 below.

```

%%Calculate the utility value for each parameter
if sex == 'm'
    TC1 = 5.2;
elseif sex == 'f'
    TC1 = 5.3;
end
if diabetic == 'y'
    BPH1 = 130; BPL1 = 80;
elseif diabetic == 'n'
    BPH1 = 140; BPL1 = 90;
end
U1 = weight1*(HB-HB1)/HB1;    U2 = weight2*(BMI-BMI1)/BMI1;    U3 =
weight3*(BPH-BPH1)/BPH1;
U4 = weight4*(HDL-HDL1)/HDL1;    U5 = weight5*(MAP-MAP1)/MAP1;    U6 =
weight6*(TC-TC1)/TC1;
str=['disp: U1='];
%display the constraint matrix
eye(6);
%calculate the optimisation values using the formulae below
f = [U1 ; U2 ; U3 ; U4 ; U5 ; U6 ]; %Objective function to be maximised
A = [ 1 0 0 0 0 0          % The constraints in a matrix format
      0 1 0 0 0 0
      0 0 1 0 0 0
      0 0 0 1 0 0
      0 0 0 0 1 0
      0 0 0 0 0 1] ;
b = [HB1; BMI1; BPH1; -HDL1; MAP1; TC1] ;    %Po which are the upper/lower
lb = zeros(6,1)
tic
[x,fval, exitflag, output, lambda] = linprog(f,A,b,[],[],lb),
lambda.ineqlin, lambda.lower
toc

```

Figure 5.6: CMAUT Source code from the MATLAB editor

Model1_GUI

Percentage Risk for Heart Diseases Model I

Patient serial no. Patient group no.

Patient age Diabetic (y/n)

Gender (m/f) Smoke (y/n)

HB ECG (y/n)

BMI

BPH Initial Clinical Percentage Risk % **19.729**

BPL

HDL

MAP

Total Choles

	X-Attributes	OPT values
1	X1 - HB	101
2	X2 - BMI	101
3	X3 - BPH/L	1
4	X4 - HDL	1
5	X5 - MAP	101
6	X6 - TC	101

Figure 5.7: Output GUI for the CMAUT framework MODEL-I

5.5.2 Determination of APR Risk using CVD CMAUT Framework Model 2

- Determination of the APR using the CVD Model2 CMAUT framework

The Model-2 of the CMAUT framework was developed using the data from the HSE, (2006) report, which was filtered to include only participants who are over 30 years old. This is because research conducted on Web CVD risk calculators, which were used to benchmark the CMAUT framework, are designed for adults between the ages of 32 to 72 years (Chuang et al. 2007). Thus model 2 was designed with the 3645 participants' data in Appendix 3.4C.

- Model 2 of the CMAUT framework

This second CMAUT model 2 was built using the Variables in the Equation obtained from statistical analysis of the HSE, (2006) data of 3645 participants who were over 30 years with full clinical data. This was done to compare the results from the Framingham equation and web calculators with the CMAUT Framework.

Example 1: Participant 2 from the HSE, (2006), has the following information:

The attributes of Participant 2, were first converted into utility unit by transforming each of the coefficients of regression (β_i) into percentage using the equation $U = (\beta_i \frac{P_i - P_o}{P_o})$.

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
42,831,101.00	No	No	35	Men	White	64.50	18.15	135.00	70.50	1.70	92.00	No	4.10	Yes	No	No

In the CMAUT model the objective function, which uses the utility units (U) and their attributes are maximised subject to the inequality constraint matrix.

The procedure in the Figure 5.8 below was used to determine the logistic Regression coefficients with the SPSS. The binary logistic regression analysis conducted gave the SPSS results in Table 5.5 and Figure 5.9.

Again, it was noted that OmsysvalBPL, which is the Low Blood pressure is not in the equation as shown in Figure 5.9, although the entire risk factors were input into SPSS for binary logistic regression analysis.

Point-and-click
 Step 1: Analyse → Regression → Binary Logistic
 Step 2: Select bp1 as Dependent variable and move it to the text box
 Step 3: Select: OmpulvalHB, BMI, OmdiavalBPL, OmsysvalBPH, Hdlval1HDL, OmmapvalMAP, CholvalTotalCholestrol and move them to Covariate text box on the right
 Step 4: Click on Options,
 Step 5: Select classification plot, Hosmer and Lemeshow Test goodness fit, correlation estimate, CI of 95%
 Step 6: Click on OK

Figure 5.8: Procedure for determining logistic Regression in SPSS

The model 2 was built using the regression coefficients obtained from the SPSS procedure discussed in section 5.5 and in Figure 5.8. The regression coefficients obtained from the SPSS as shown in Figure 5.9 are the weight attached to each attribute. This is applied in the utility function formula (equation 5.1) to calculate the utility unit (U) of each specific attribute and the overall utility unit is the sum of all the individual utility units.

Table 5.5: Classification Table 2(a,b) Total participants = 3645

Observed			Predicted		Percentage Correct
			bp1		
			0	1	
Step 0	bp1	0	2968	0	100.0
		1	677	0	.0
Overall Percentage					81.4

a Constant is included in the model.

b The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	OmpulvalHB	.174	.156	1.240	1	.265	1.190
	BMI	.076	.010	56.237	1	.000	1.079
	OmsysvalBPH	-.229	.234	.958	1	.328	.795
	Hdlval1HDL	.216	.123	3.073	1	.080	1.241
	OmmapvalMAP	.279	.234	1.416	1	.234	1.321
	CholvalTotalCholestrol	.058	.043	1.804	1	.179	1.060
	Constant	-10.076	.524	370.297	1	.000	.000

a Variable(s) entered on step 1: OmpulvalHB, BMI, OmsysvalBPH, Hdlval1HDL, OmmapvalMAP, CholvalTotalCholestrol.

Figure 5.9: Output from the logistic Regression in SPSS

In the CMAUT model 2, the attributes in the expression below are converted into utility units using the weight percentage equation $U = wt \frac{P_i - P_0}{P_0}$.

$$G_1: [(H_1 \wedge K_2 \wedge B_3), R_1, I_2, B_{H3}, B_{L4}, H_5, M_6, T_7] \quad (5.13)$$

Since the simulation of the framework model 2 is based on the data from HSE,(2006), the individual weights are from the binary logistic regression based on SPSS where $y = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$. The coefficient of regression (β_i) of each attribute in the equation after the SPSS analysis of over 30 years old participants' data are shown in Figure 5.9 above. The weights of attributes are OmpulvalHB = 0.174, BMI = 0.076, OmsysvalBPH = -0.229, Hdlval1HDL = 0.216, OmmapvalMAP = 0.279, CholvalTotalCholestrol = 0.058 and the constant value = -10.076. The number of attributes in the equation is reduced to six since the OmdymaBPL does not appear in the SPSS results.

Therefore the above equation (5.13) can be written as follows:

$$y = -10.076 + .174x_1 + .076x_2 + -.229x_3 + .216x_4 + .279x_5 + .058x_6. \quad (5.13)$$

- Conversion of the attributes into utility unit

For the conversion of attributes into utility unit the first procedure is that each of the coefficients of regression (β_i) is converted into utility function using the equation (β_i) into percentage using the equation $U = (\beta_i \frac{P_i - P_0}{P_0})$. Therefore the utility unit for each attribute is calculated as follows:

$\begin{aligned} \text{UHB} = \text{UR} &= (\beta_i (R_1 - R_0)/R_0) = 0.174 ((50-34)/50) = 5.568 \\ \text{UBMI} = \text{UV} &= ((\beta_i (V_1 - V_0)/V_0)) = 0.076 ((25.5-13.2)/25.5) = 3.665 \\ \text{UBPH} = \text{UP} &= ((\beta_i (P_{H1} - P_{H0})/P_0)) = -0.229 ((140-122.5)/140) = -2.86 \\ \text{UHDL} = \text{UD} &= ((\beta_i (D_1 - D_0)/D_0)) = 0.216 ((1.2-1.8)/1.2) = -10.8 \\ \text{UMAP} = \text{UM} &= ((\beta_i (M_1 - M_0)/M_0)) = 0.279((100 -100)/100) = 0 \\ \text{UTC} = \text{UT} &= ((\beta_i (T_1 - T_0)/T_0)) = 0.058 ((5-5.2)/5) = -0.232 \end{aligned}$
--

Box 5.5: Computation of the Utility Units

From the utility unit calculation, the expression (5.13) can be rewritten as

$$G_1: [(H_1 \wedge K_2 \wedge B_3), U_1, U_2, U_{H3}, U_{L4}, U_5, U_6,] \quad (5.14)$$

From the calculated utility unit the expression (5.14) can be rewritten as

$$G_1: [(H_1 \wedge K_2 \wedge B_3), U_R, U_V, U_{PH}, U_D, U_M, U_T,] \quad (5.15)$$

The expression (5.15) shows the relative effect and influence each attribute has on the disease under consideration. The calculated U_i vector values in expression (5.14) indicate the relative strength of each of the attribute in the disease domain and must be converted into a formal expression that can be manipulated and analysed for decision making (Deutsch et al., 1994). The calculated utility unit U_i , of each attribute is a vector value, which is attached to the respective attribute variables in the format $U_n X_n$. Since the organs in the disease domain complement each other the overall utility unit is the arithmetical sum of the individual utility units. Therefore the above expression (5.15) and the calculated individual utility units with their respective attributes can be written as $Z = \sum_i^n (U_1 X_1 + U_2 X_2 + \dots + U_n X_n)$. This expression is formulated as the objective function that must be maximised and must include the constant as well as the beta values from the SPSS regression. In generic terms the objective function for both model 1 and 2 are expressed as follows:

$$Z = \sum_i^n (a + U_R X_R + U_V X_V + U_{PH} X_{PH} + U_D X_D + U_M X_M + U_T X_T) \quad (5.16)$$

When the individual calculated utility units for model 2 are substituted into the objective function, the final function is as follows:

$$Z = \sum_i^n ((-10.076) + 5.568 X_R + 3.665 X_V + (-2.86) X_{PH} + (-10.8) X_D + (0) X_M + (-0.232) X_T) \quad (5.17)$$

The objective function in expression (5.17) is maximised subject to set of inequality constraints matrix. In this set of inequality constraint matrix, each inequality stands for the measured attribute that must be less than or equal to the acceptable value, which meets the CVD standard risk factors stated in NICE, (2006) guidelines.

$$\begin{array}{rcl}
CX_R + 0 & + & 0 + 0 + 0 + 0 + 0 \leq 50 \\
0 + CX_B + & & 0 + 0 + 0 + 0 + 0 \leq 25.5 \\
0 + & 0 + & CX_{PH} + 0 + 0 + 0 + 0 \leq 140 \\
0 & + & 0 + 0 + CX_H + 0 + 0 \geq 1.2 \\
0 & + & 0 + 0 + 0 + 0 + CX_M + 0 \leq 100 \\
0 & + & 0 + 0 + 0 + 0 + 0 + CX_T \leq 5
\end{array}$$

Box 5.6: Set of inequalities for the constraints matrix

To solve the above optimisation problem using the model 2 results the LP LINPROG method in MATLAB is used. First all the coefficients in the objective function are formatted as in Box 5.7 below:

$$F = -10.076 + 5.568 + 3.665 - 2.86 - 10.8 + 0 + -232$$

A = the coefficient in the unit constraint matrix; see below

$$b = [50, 25.5, 140, -1.2, 100, 5];$$

$$1b = \text{zeros}(6, 1);$$

Secondly, the program calls a linear programming routine in MATLAB library.

$$[X, fval, \text{exitflag}, \text{output}, \text{lamba}] = \text{Linprog}(f, A, b, [], [], 1b);$$

Box 5.7: The CVD Optimisation algorithm using linear programming in MATLAB

In Box 5.6, the set of inequalities **A**, in the linear programming function is expressed using identity matrix or unit matrix format as shown below

$$A = \begin{array}{cccccc}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{array}$$

Box 5.8: Set of Inequalities unit matrix A used for Optimisation in MATLAB

The unit matrix in Box 5.6 and Box 5.8 were used to develop the MATLAB CVD Optimisation algorithm shown in Box 5.7. The full Optimisation algorithm was written using MATLAB source code shown in Figure 5.10 below. The results of the CVD risk after executing the MATLAB Optimisation program is shown in Figure 5.11 below.

```

%%Calculate the utility value for each parameter
if sex == 'm'
    TC1 = 5.2;
elseif sex == 'f'
    TC1 = 5.3;
end
if diabetic == 'y'
    BPH1 = 130; BPL1 = 80;
elseif diabetic == 'n'
    BPH1 = 140; BPL1 = 90;
end
U1 = weight1*(HB-HB1)/HB1;    U2 = weight2*(BMI-BMI1)/BMI1;    U3 =
weight3*(BPH-BPH1)/BPH1;
U4 = weight4*(HDL-HDL1)/HDL1;    U5 = weight5*(MAP-MAP1)/MAP1;    U6 =
weight6*(TC-TC1)/TC1;
str=['disp: U1='];
%display the constraint matrix
eye(6);
%calculate the optimisation values using the formulae below
f = [U1 ; U2 ; U3 ; U4 ; U5 ; U6 ]; %Objective function to be maximised
A = [ 1 0 0 0 0 0          % The constraints in a matrix format
      0 1 0 0 0 0
      0 0 1 0 0 0
      0 0 0 1 0 0
      0 0 0 0 1 0
      0 0 0 0 0 1] ;
b = [HB1; BMI1; BPH1; -HDL1; MAP1; TC1] ;    %Po which are the upper/lower
lb = zeros(6,1)
tic
[x,fval, exitflag, output, lambda] = linprog(f,A,b,[],[],lb),
lambda.ineqlin, lambda.lower
toc

```

Figure 5.10: Source code for CVD CMAUT MODEL – II

Percentage Risk for Heart Diseases Model II

Patient serial no. Patient group no.

Patient age Diabetic (y/n)

Gender (m/f) Smoke (y/n)

HB ECG (y/n)

BMI

BPH Initial Clinical Percentage Risk % **18.0074**

BPL

HDL

MAP

Total Choles

	X-Attributes	OPT values
1	X1 - HB	50
2	X2 - BMI	26
3	X3 - BPHL	1
4	X4 - HDL	1
5	X5 - MAP	100
6	X6 - TC	6

Figure 5.11 Output screen from CVD CMAUT MODEL – II

5.6 Validation of CMAUT framework using Prevalence and Kappa statistic

To validate the two CMAUT models, Prevalence and Kappa criteria, which indicate the percentage agreement between hypertensive NO and YES status of the participants were used. The kappa values indicate the degree of agreement between the results of the APR from the CMAUT models and the results of the GP identified hypertensive NO and YES status. The diagnosis risk results were validated using the results of the GP identified hypertension, which are marked (1) for YES and non-hypertension marked (0) for NO as compared to the NICE, (2006) recommendation of 20%.

5.6.1 Prevalence Computation

In this research, prevalence is the extent to which hypertensive disease can be found in the population of participants who took part in the HSE 2006 survey (Sanderson et al. 2006). This research compares and contrasts the prevalence value from the HSE, (2006) and the results from the two CMAUT diagnosis models. To compute the prevalence value the formula for $\text{Prevalence} = a / (a+b)$, which was explained in Chapter 3 and used in this chapter (Campbell et al., 2007).

- Actual computed Prevalence from the (HSE, 2006) report:

The total population of participants who took part in the survey were 21,399. As discussed in section 5. 5 above 4165 people who were over 16 years old and provided complete information are used for model 1. The 3645 participants who were over 30 years old and submitted all the CVD data as well as diagnosed by the GP were used for model 2.

To determine the number of people in the HSE population, who were identified as been hypertensive by the GP a statistical analysis was conducted. In the report, the participants who were diagnosed by GP that they had hypertension were marked YES or (1) and non-hypertension participants were marked NO or (0) in the spread sheet and SPSS respectively.

From the Excel spread sheet that contains the 3645 participants, using Advance filter tool, it was identified that 677 people were diagnosed as been hypertensive YES while 2968 were hypertension NO. The prevalence of GP identified hypertensive from the report also known as actual prevalence was computed as follows:

$$\text{Actual Prevalence} = GP \text{ bpyes} / (GP \text{ bpyes} + GP \text{ bpno}) \quad (5.18)$$

$$= (677 / (677 + 2968)) = 0.1857. \quad \text{Therefore the Actual Prevalence} = 0.1857$$

- Prevalence value for CMAUT Model 1:

As discussed above in section 5.5, model 1 was designed using participants who are over 16 years old and have who fulfilled all the inclusive requirements. After building the prototype, it was used to determine the percentage clinical risk for each of the 4165 participants. For the purpose of comparison only participants were over 30 years old were used for the analysis in order to prevent any prejudice.

According to NICE, (2006), the CVD percentage risk threshold value of a participant been hypertensive YES must be 20% or higher. Whereas any person that has CVD percentage risk value of less than 20%, must be declared as hypertension NO. Based on this concept, the percentage risk value of each participant was computed and grouped as follows:

- If clinical percentage risk value is $\geq 20\%$ then the participants is hypertension YES
else when the percentage risk value is $< 20\%$ then the participant is hypertension NO.

The condition and computation of the Prevalence value for CMAUT Model 1 was programmed in MS Excel software and the results are in Box 5.9 below:

Number of BPYes (based on 20% threshold) = 43
 Number of BPNo(based on 20% threshold) = 2714
 Percentage agreement between model 1 and Actual
 Percentage agreement of YES between model 1 and Actual GP YES
 $\text{Agreed Yes (model1)/Actual YES} * 100 \% = (43 / 677) * 100 = 6.35\%$
 Percentage agreement of NO between model 1 and Actual GP NO
 $\text{Agreed No (model1)/Actual No} * 100 \% = (2714 / 2968) * 100 = 91.44\%$

Box 5.9: Computation of Prevalence value for CMAUT Model 1

Therefore using the expression (5.18) and converting the equation to match the CMAUT framework model 1 the following is obtained:

The Prevalence value for Model 1 = $\text{Model1 bpyes} / (\text{Model1 bpyes} + \text{Model1 bpno})$
 $= (43 / (43 + 2714)) = 0.1857$. Therefore the Prevalence for CMAUT model1 = 0.1857

- Prevalence for the CMAUT Model 2:

As discussed above, model 2 was designed using the data of the participants who are over of 30 years and fulfilled all the inclusive requirements. The prototype built was used to determine the APR value for each of the 3546 participants. For the purpose of comparison only participants who are over 30 years old were used for the analysis to prevent prejudice.

Again based on the (NICE, 2006) recommendation, the threshold percentage risk value of a participant been hypertension YES must be 20% or higher. When a participant has a CVD percentage risk value of less than 20%, they must be declared hypertension NO. Based on this concept, the percentage risk of each participant was computed and grouped as follows:

- If clinical percentage risk value $\geq 20\%$ then the participants is hypertension YES
 else when the percentage risk value is $< 20\%$ then the participant is hypertension NO.

The condition and computation of the Prevalence value for CMAUT Model 2 was programmed in MS Excel software and the results are in Box 5.10 below:

Number of BPYes (based on 20% threshold) = 30
 Number of BPNo (based on 20% threshold) = 2889
 Percentage agreement between model 2 and Actual
 Percentage agreement of YES between model 1 and Actual GP YES
 $\text{Agreed Yes (model1) / Actual YES} * 100 \% = (30 / 677) * 100 = 4.4313\%$
 Percentage agreement of NO between model 1 and Actual GP NO
 $\text{Agreed No (model1) / Actual No} * 100 \% = (2889 / 2968) * 100 = 97.3383\%$

Box 5.10: Computation of Prevalence value for CMAUT Model 2

- Prevalence for Model 2

Using the expression (5.18) above and converting it to match model 2 the following is obtained. Prevalence for Model 2 = Model2bpyes/ (Model2bpyes + Model2bpno)
 $= (30 / (30+2889)) = 0.0103$. Therefore the computed Prevalence for model2 is 0.0103.

5.6.2 Computation of Kappa statistic

Kappa is the measure of a prediction model's ability to determine the degree of the agreement between two predictors (Campbell et al. 2007). According to Viera, (2005), Kappa is a method used to determine the degree of agreement between two predictors using quantitative measure. In this research, kappa statistic is use to determine the degree of agreement of the actual results from the HSE survey and the results from the two CMAUTF models.

To compute the kappa value the number of participants with and without hypertension as diagnosed by the GPs as well as the resultant values from the CMAUT models were presented in a tabular format as shown in Table 3.2. Using Table 3.2, the kappa value k is computed using the formula $k = \frac{P_o - P_e}{1 - P_e}$ where $P_e = [(n_1/n) * (m_1/n) + ((n_0/n) * (m_0/n))]$ and $P_o = [(a + d)/N]$ (Viera, 2005) (Cunningham et al, 2009).

- Kappa statistics for MATLAB Model I

As discussed in section 5.6, according to NICE, (2006) the threshold value of a participant been hypertensive is equal or greater than 20%. Based on the 20% threshold the number of participants predicted hypertension YES are those that have percentage risk value of over 20% was 43 and participant with less than 20% risk was 2968. From the survey report, the results revealed that the diagnosed hypertension YES was 677 and hypertension NO was 2714. These values are entered into the Table 3.2 and presented in Table 5.6 below.

Table 5.6

	Yes	No	Total
Yes	677	2968	3645
No	43	2714	2757
	720	5682	6402

- Computation of Kappa value for model 1

Using the information from Table 5.6 and Kappa formula discussed above the following steps are used to calculate the Kappa (k) for model 1:

$$k = \frac{P_o - P_e}{1 - P_e}$$

$$P_o = [(677 + 2714)/6402]$$

$$P_o = 0.5297$$

$$P_e = [(n_1/n) * (m_1/n) + ((n_0/n) * (m_0/n))]$$

$$P_e = [(3645/6402) * (720/6402) + (5682/6402) * (2757/6402)]$$

$$P_e = 0.4462$$

$$k = (0.5297 - 0.4462)/(1 - 0.4462)$$

$$k = 0.1508$$

Box 5.11: Computation of Kappa value for CMAUT Model 1

From the Kappa Interpretation Table in Viera, (2005) and Cunningham et al, (2009) the $k = 0.1508$ means the models slightly agree with each other. This means the APR values from CMAUT model 1 slightly agree with the actual GP diagnosed hypertension YES and NO.

- Kappa statistics for MATLAB Model II

The same method was used to determine the kappa for model 2 using the NICE recommended percentage risk threshold value of 20%. For model 2, the number of predicted hypertension YES that is the percentage risk, which is more than 20%, was 30 and those less than 20% was 2889. From the survey results, it is revealed that for the diagnoses, the hypertension YES was 677 and hypertensive NO was 2714. These values are entered into the Table 3.2 and presented in Table 5.7 below.

Table 5.7

	Yes	No	Total
Yes	677	2968	3645
No	30	2889	2919
	707	5857	6564

- Computation of Kappa value for model 2

Using the information from Table 5.7 and the kappa formula discussed above the following steps are used to calculate the kappa value for model 2:

$$\begin{aligned}
 &= [(OA - \text{expected agreement}) / (100\% - \text{expected agreement})] \\
 &= [((677+2889)/50 - 0.50) / (1 - 0.50)] \\
 &= [(0.71-0.50)/(1-0.50)] \\
 &= 0.42
 \end{aligned}$$

Box 5.12: Computation of Kappa value for CMAUT Model 2

Using the same Kappa Interpretation Table in Viera, (2005) the k value of 0.42 means the results from the CMAUT model 2 fairly agrees with the percentage risks results from the actual GP diagnosed hypertension YES and NO.

5.7 Simulation Results for CMAUT Diagnosis Framework model 1 and 2

5.7.1 Simulation Results, Tables and Figures for CMAUT Diagnosis model 1

The Tables and Figures in this section are the results of inputting the demographic and clinical data of each of the selected 3654 participants into the CVD CMAUT Diagnosis framework model 1.

The Table 5.8A is the raw data of the first 10 participants from the list of 3654 participants that is repeated for reference purposes. Table 5.8B, which is at the end of this Thesis contains the data of the first 30 participants and the data of the entire group is in Appendix Table 5.8C in electronic format. This is followed by Table 5.9A, which contains the results of the calculated APR values and the variable attribute values of each of the first 10 participants

from Model I 3645 data sets. The Table 5.9B, at the end of the Thesis contains APR values of the first 30 participants and the results of the entire group are in Appendix Table 5.9C in electronic format.

To evaluate the CMAUT CVD Diagnosis framework models 1 and 2, the criteria metrics True Positive Rate (TPR), False Positive Rate (FPR), Positive Likelihood ratio (LRP) and Negative Likelihood ratio (LRN) were calculated and presented in Table 5.3. Table 5.10A shows the results of the computation of TPR, FPR, LRP and LRN for the CMAUT CVD Diagnosis model 1 using the APR values for the first 10 participants from the 3645 data set. The Table 5.10B, at the end of the Thesis contains results of the first 30 participants and the results of the entire group are in Appendix Table 5.10C in electronic format.

MATLAB Model I for 3645 participants in the category of over 16 years old.

Table 5.8A: Raw data of the first 10 participants

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
71,831,101.00	No	No	66	Women	White	89.00	14.32	159.00	70.00	1.90	99.50	No	6.90	No	No	Yes
34,031,101.00	No	No	84	Women	White	48.50	16.17	112.00	63.50	2.20	80.00	No	5.00	Yes	Yes	Yes
72,604,102.00	No	No	59	Women	White	36.50	16.19	109.50	73.00	2.00	85.00	No	6.00	No	No	No
13,008,101.00	Yes	Yes	50	Women	White	43.00	16.65	117.00	74.00	1.70	88.50	No	6.00	Yes	Yes	Yes
39,139,101.00	No	No	34	Women	White	48.00	16.81	102.00	54.00	1.80	70.00	No	6.50	Yes	No	No
47,856,102.00	No	No	51	Women	White	44.00	16.85	100.50	56.50	1.90	71.00	No	5.10	No	No	No
37,710,101.00	No	No	61	Women	White	43.00	17.43	120.00	77.00	1.20	91.50	No	5.50	No	No	No

Table 5.9A: Absolute percentage risks and variable attributes values for the first 10 participants

Pserial no.	Grp	Bp1	Age	Sex	X1	X2	X3	X4	X5	X6	%PR
13,956,102.00	No	No	60	Women	50.0	25.5	0.02	0.00	50.00	5.22	-14.6
63,535,102.00	Yes	Yes	30	Women	45.8	25.9	4.15	0.00	100.19	4.28	-15.4
71,831,101.00	No	No	66	Women	0.0	25.5	140.00	0.00	100.00	0.00	-16.7
34,031,101.00	No	No	84	Women	50.0	25.5	0.00	0.00	100.00	5.30	-18.0
72,604,102.00	No	No	59	Women	50.0	25.5	0.00	0.00	100.00	0.00	-18.9
13,008,101.00	Yes	Yes	50	Women	50.0	25.5	0.00	0.00	100.00	0.00	-16.3
39,139,101.00	No	No	34	Women	50.0	25.5	0.00	0.00	100.00	0.00	-21.4
47,856,102.00	No	No	51	Women	50.0	25.5	0.01	0.00	100.00	5.23	-21.9
37,710,101.00	No	No	61	Women	50.0	25.5	0.00	0.00	100.00	0.00	-15.2
54,256,101.00	No	No	31	Women	50.0	25.5	0.00	0.00	100.00	5.30	-14.1

Table 5.10A: Calculation of TPR, FPR, LRP, and LRN, for the MATLAB Model I for the first 10 participants

Pserial no.	Grp	BpI	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRM
13,956,102.00	No	No	60	Women	-14.6	0	1	1	0.9997	3574	0
63,535,102.00	Yes	Yes	30	Women	-15.4	0	1	1	0.9994	1787	0
71,831,101.00	No	No	66	Women	-16.7	0	1	1	0.9991	1191.33	0
34,031,101.00	No	No	84	Women	-18.0	0	1	1	0.9988	893.5	0
72,604,102.00	No	No	59	Women	-18.9	0	1	1	0.9985	714.8	0
13,008,101.00	Yes	Yes	50	Women	-16.3	0	1	1	0.9982	595.666	0
39,139,101.00	No	No	34	Women	-21.4	1	0	0.9966	0.9982	510.571	0
47,856,102.00	No	No	51	Women	-21.9	1	0	0.9932	0.9982	446.75	0
37,710,101.00	No	No	61	Women	-15.2	0	1	0.9932	0.9979	397.11	0
54,256,101.00	No	No	31	Women	-14.1	0	1	0.9932	0.9976	357.4	0

Figure 5.12 shows the graphical representation of the computed APR values of each of the 3645 participants against their individual Participant Identification Number (PIND). This data used for plotting the graph is from Table 5.8C.

Figure 5.13 is the prediction accuracy graph of the computed TPR value against the FPR value of each of the 3645 participants. This data used for plotting this graph is from Table 5.9C. From the ROC graph, the Area under the Curve (AUC) for model 1 is calculated, by the summation of all the APR data points using the trapezoidal method. The actual area is obtained by subtracting the sum of all the APR data points from the sum of all the diagonal reference data points. See details in chapter 3 and 9.

Figure 5.14 shows the graph of the discriminatory ability of the Diagnosis model 1. This was constructed by first computing the sensitivity (aka TPR) and selectivity (aka FPR) of each of the 3645 participants. See Table 7.10C for the results of the computation of the TPR and FPR values for model 1. A graph of the sensitivity and selectivity are plotted against the recommended criterion. In this research NICE, (2006) the recommended criterion of 20% is used while the interception and the degree of accuracy are discussed in chapter 9.

Figure 5.15 shows the graph of the performance accuracy of the APR of each participant's value from the Diagnosis framework model 1. The procedure used to calculate the value of the positive Likelihood ratio $LR+ = (TPR/1-TNR)$ and the negative Likelihood ratio $LR- = (1-TPR/TNR)$ for all the 3645 participants. See Table 5.10C for the results of the computation of the positive Likelihood ratio $LR+$ and the negative Likelihood ratio $LR-$ for model 1.

All the positive LR and negative LR values were plotted on the Y-axis and the PIND of each participant on the X-axis. The graph is shown in Figure 5. 15 and it is discussed in chapter 9.

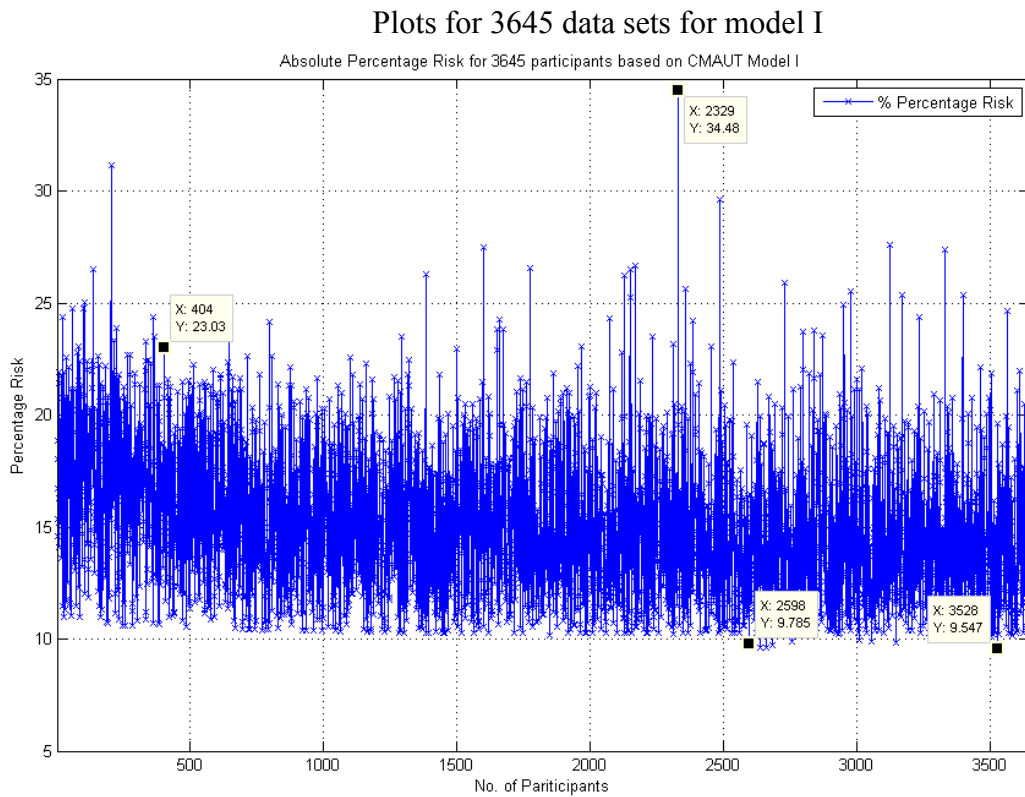


Figure 5.12 Absolute Percentage Risk for CMAUT Model – I

ROC and AUC

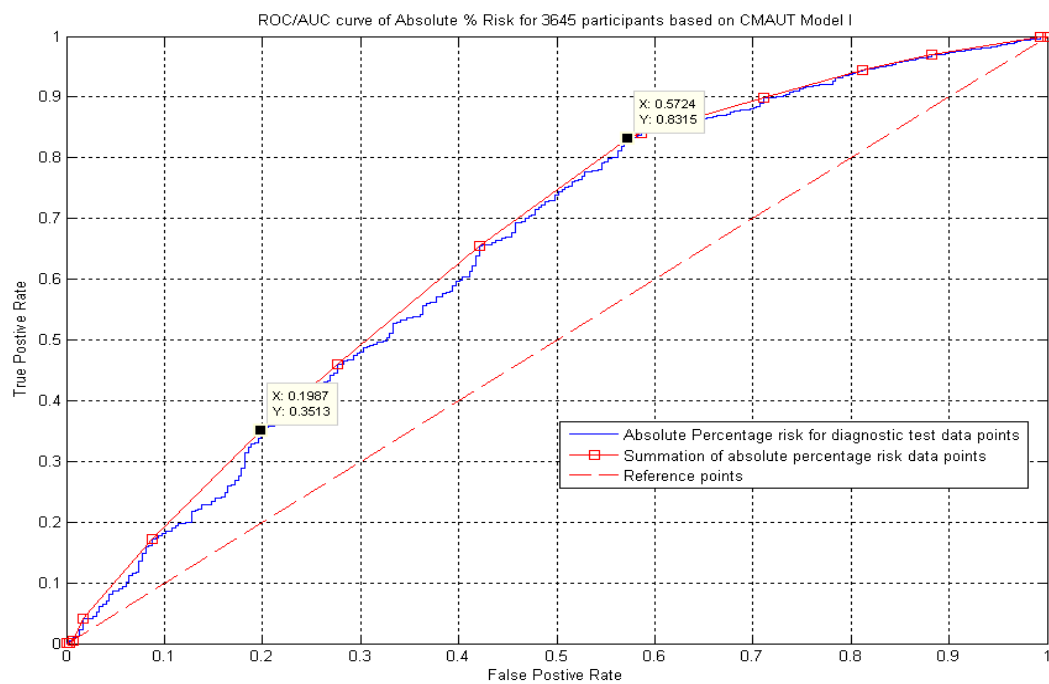


Figure 5.13: ROC/AUC curve of absolute percentage risk Model - I

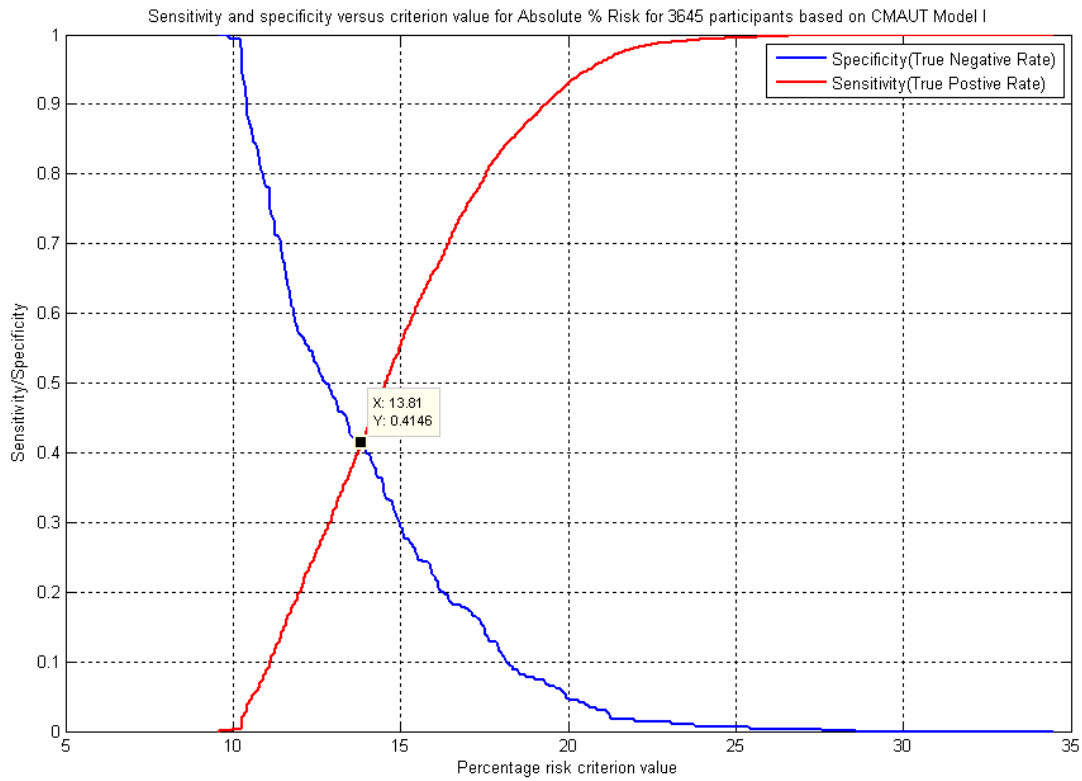


Figure 5.14: Sensitivity and specificity for Model - I

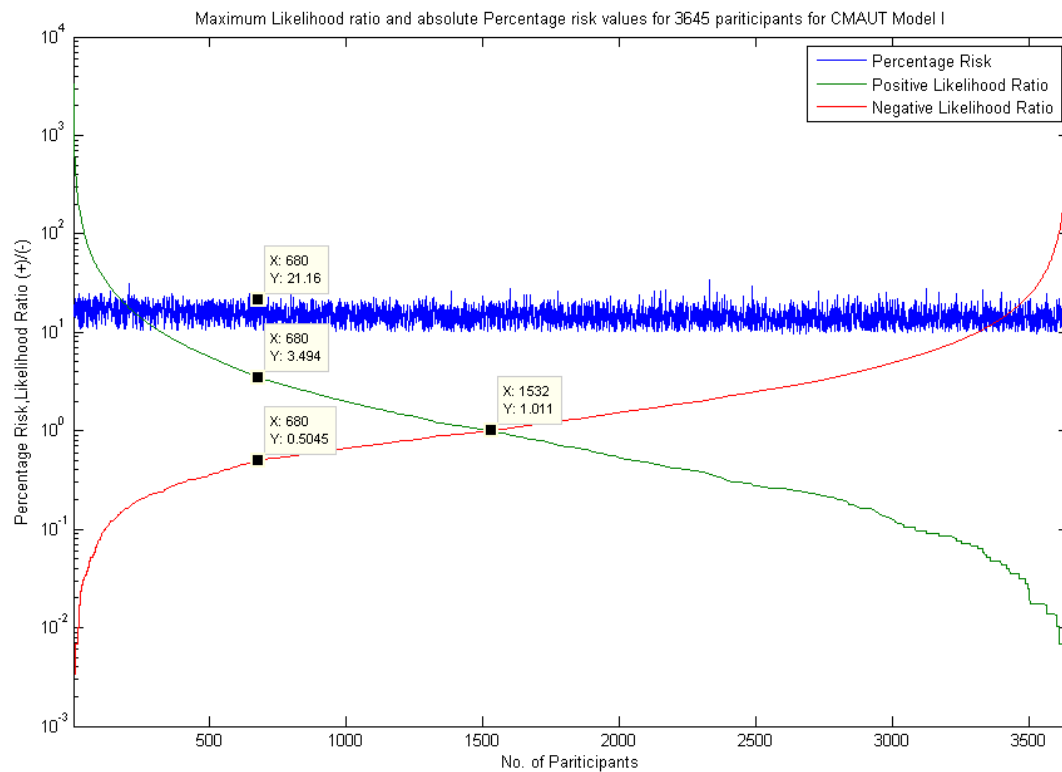


Figure 5.15: Maximum likelihood ratio model - I

5.7.2 Simulation Results, Tables and Figures for CMAUT Diagnosis model 2

The Tables and Figures in this section are the results of inputting the demographic and clinical data of each of the selected 3654 participants into the CVD CMAUT Diagnosis framework model 2. The Table 5.9A, contains the results of the calculated APR values and the variable attribute values of each of the first 10 participants from Model II 3645 data sets. The Table 5.9B, at the end of this Thesis contains the first 30 participants and the APR results of the entire group are in Appendix Table 5.9C in electronic format.

To evaluate the CMAUT CVD Diagnosis framework models, the criteria metrics TPR, FPR, LRP and LRN were calculated and presented in Table 5.10. Table 5.10A shows the results of the computation of TPR, FPR, LRP and LRN for the CMAUT CVD Diagnosis model 2 using the APR for the first 10 participants from the 3645 data set. The Table 5.10B, at the end of this Thesis contains the results of the first 30 participants and the results of the entire group are in Appendix Table 5.10C in electronic format.

- CMAUT Model II for 3645 participants in the category of over 30 years old

Table 5.11A: Initial absolute percentage risks and attributes variable values for the first 10 participants Model II

Pserial no.	Grp	Bp1	Age	Sex	X1	X2	X3	X4	X5	X6	%PR
13,956,102.00	No	No	60	Women	50.1	22.6	31.4	1.0	50.3	3.3	12.8
63,535,102.00	Yes	Yes	30	Women	45.1	26.3	4.0	1.0	99.5	4.1	14.4
71,831,101.00	No	No	66	Women	10.3	14.1	102.7	1.0	19.2	6.5	12.2
34,031,101.00	No	No	84	Women	30.8	19.4	30.2	1.0	88.8	3.4	14.4
72,604,102.00	No	No	59	Women	35.7	9.1	34.4	1.0	76.7	6.5	13.4
13,008,101.00	Yes	Yes	50	Women	41.6	21.7	25.4	1.0	89.3	2.9	13.5
39,139,101.00	No	No	34	Women	36.1	21.3	2.4	1.0	100.2	2.7	19.1
47,856,102.00	No	No	51	Women	41.9	21.4	3.8	1.0	100.1	3.4	19.4
37,710,101.00	No	No	61	Women	48.9	24.4	3.4	1.0	98.7	5.4	14.1
54,256,101.00	No	No	31	Women	42.0	19.7	33.6	1.0	68.1	3.9	11.7

Table 5.12A: Calculation of TPR, FPR, LRP, and LRN, for the MATLAB Model II for the first 10 participants

Pserial no.	Grp	Bp1	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRM
13,956,102.00	No	No	60	Women	12.8	0	1	1	1.000	3571.43	0
63,535,102.00	Yes	Yes	30	Women	14.4	0	1	1	0.999	1785.71	0
71,831,101.00	No	No	66	Women	12.2	0	1	1	0.999	1191.90	0
34,031,101.00	No	No	84	Women	14.4	0	1	1	0.999	893.66	0
72,604,102.00	No	No	59	Women	13.4	0	1	1	0.999	714.80	0
13,008,101.00	Yes	Yes	50	Women	13.5	0	1	1	0.998	595.59	0
39,139,101.00	No	No	34	Women	19.1	0	1	1	0.998	510.46	0
47,856,102.00	No	No	51	Women	19.4	0	1	1	0.998	446.83	0
37,710,101.00	No	No	61	Women	14.1	0	1	1	0.997	397.14	0
54,256,101.00	No	No	31	Women	11.7	0	1	1	0.997	357.40	0

Figure 5.16 shows the graph of the computed APR value for each of the 3645 participants against their individual PIND. This data used for plotting the graph is from Table 5.11C.

Figure 5.17 is the prediction accuracy graph of the computed TPR against the FPR values of each of the 3645 participants. The data used for plotting this graph is from Table 5.11C. From the ROC graph, the AUC for model 2 is calculated, by using the summation of all the APR data points and applying the trapezoidal method. The area is obtained by subtracting the sum of all the APR data points from the sum of all the diagonal reference data points. See details in chapter 3 and 9.

Figure 5.18 show the graph of the discriminatory ability of the Diagnosis model 2. This was constructed by calculating the sensitivity (TPR) and selectivity (FPR) values for each of the 3645 participants. See Table 7.12C for the results of the calculation of TPR and FPR for model 2. A graph of the sensitivity and selectivity are plotted against the NICE, (2006) recommended criterion of 20% and the interception and the degree of accuracy are discussed in the evaluation chapter 9.

Figure 5.19 shows the graph of the performance accuracy of the APR for each participant from the Diagnosis framework model 2. The formulae used for the computation are $LR+ = (TPR/1-TNR)$ and the $LR- = (1-TPR/TNR)$. Table 5.12C shows the result of the calculation of the positive and negative Likelihood ratio for all the 3645 participants using the model 2. The positive and negative LR values were plotted on the Y-axis and the PIND of each participant are on the X-axis. The graph is in Figure 5.19 and it is discussed in chapter 9.

- Plots for 3645 data sets for model II

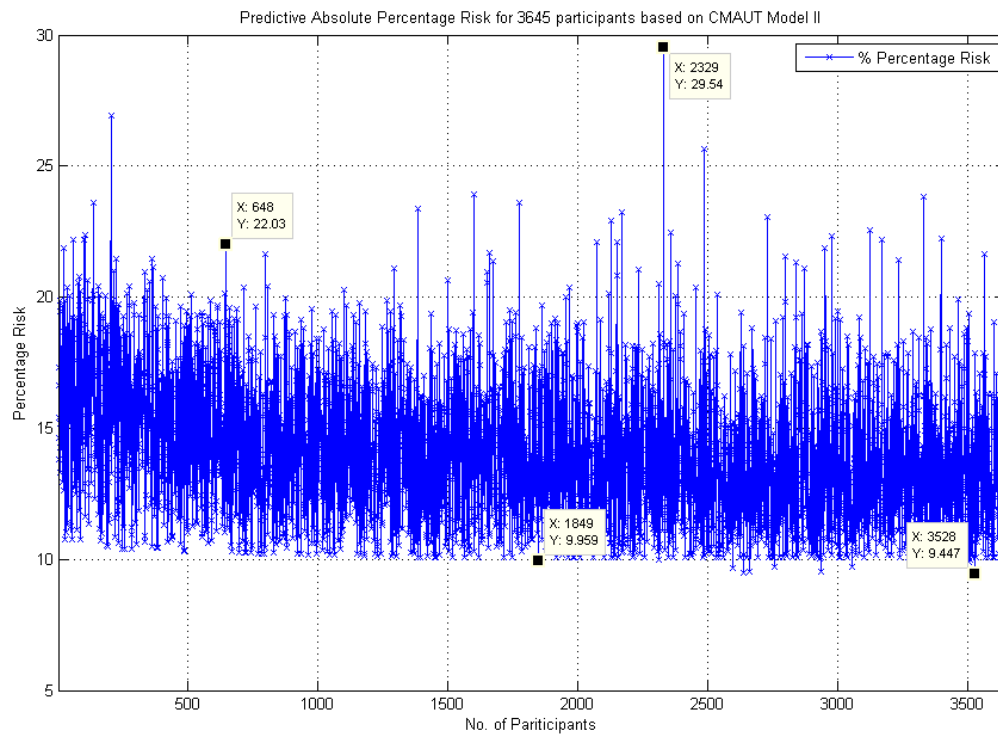


Figure 5.16: Absolute Percentage Risk Model – II

ROC and AUC

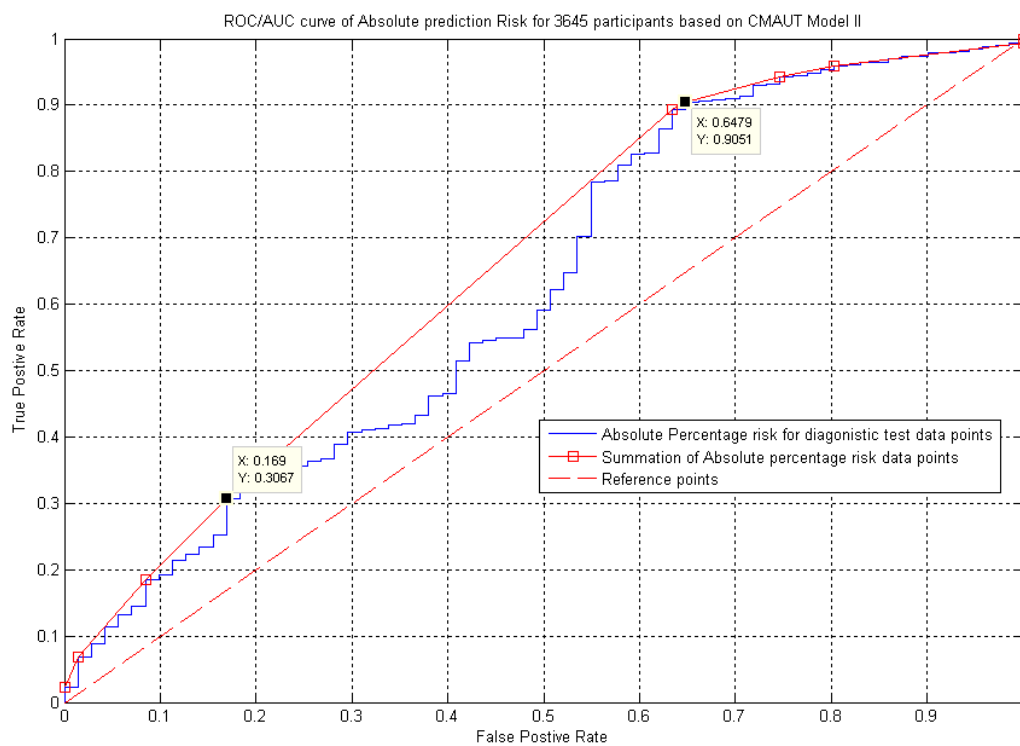


Figure 5.17: ROC/AUC curve of Absolute Predication Risk Model – II

Sensitivity and specificity

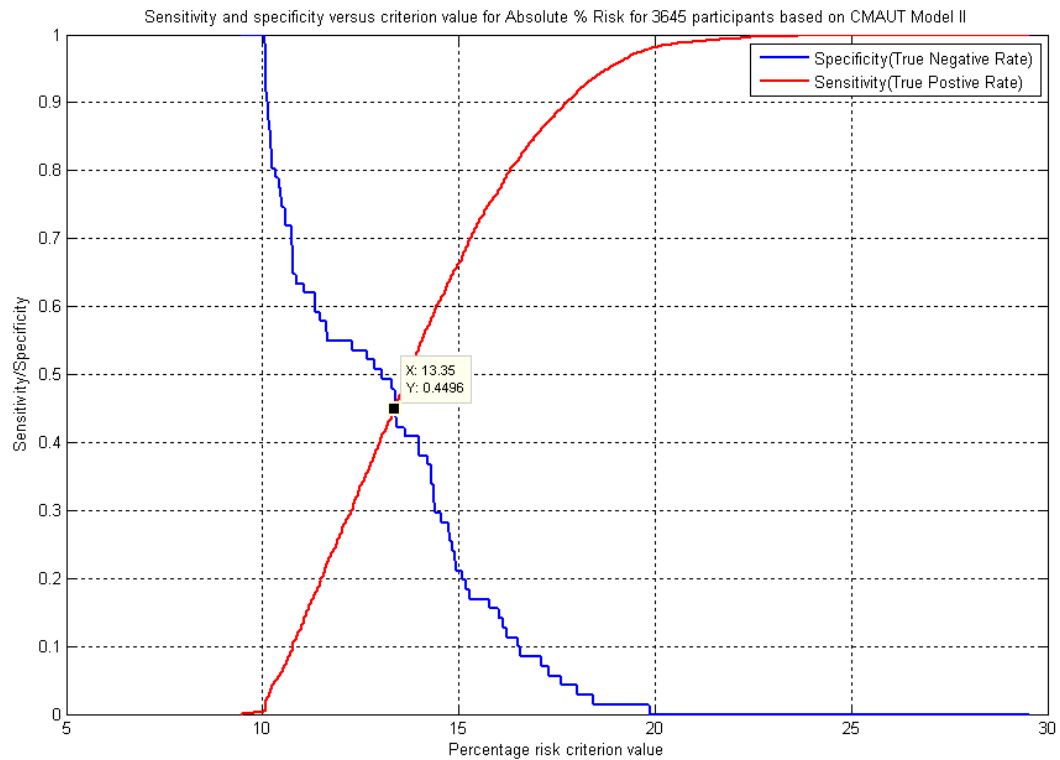


Figure 5.18: Specificity and Sensitivity versus criterion for Model - II

Likelihood ratio

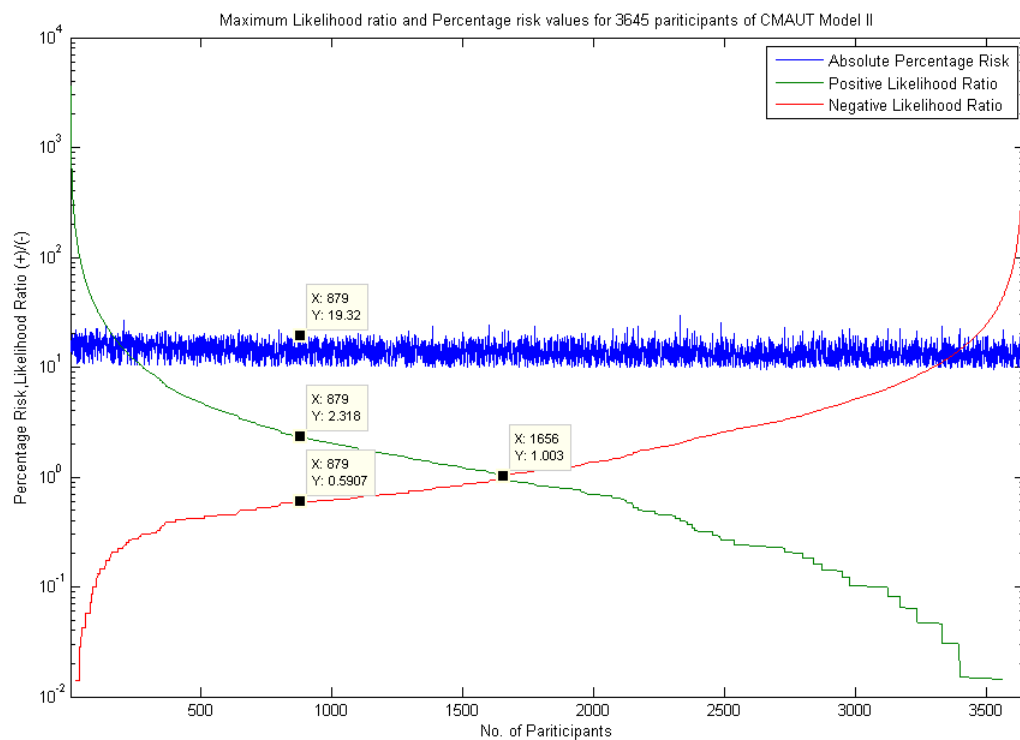


Figure 5.19: Maximum Likelihood ratio and percentage risk for Model – II

5.8 Summary:

Inferring from the calculation in section 5.6, it is deducted that the kappa for model 2, which is 0.42 is better than that of model 1 of 0.15. These values are further discussed in the evaluation chapter 9, however it must be stated that based on Viera, (2005) and Cunningham et al. (2009), model 2 is better than model 1, which is slightly agreeable as compared to model 2, which is fairly agreeable. Therefore, this research will use model 2 to implement the prototype, which will be discussed in chapter 9. According to Cook, (2007), sensitivity, and specificity analyses as well as (ROC/AUC) must be computed for each model. These analyses were carried out and they are discussed and evaluated in Chapter 9.

Incidence analysis is not conducted in this research because it is not one of the objectives; it is therefore recommended for further work. A provisional exercise was conducted where a prototype CMAUT Framework model 2 was implemented and uploaded to Amazon Cloud. The URL is http://ec2-79-125-90-51.eu-west-1.compute.amazonaws.com/NHS2_1/. This is also recommended for further work.

Chapter 6: Space Complexity and Clinical Data Reduction in CMAUT Framework

6.0 Introduction

Chapter 6 discusses the space complexity and data size reduction in the CIS when using the CMAUT Optimisation Framework. Mathematical operations were used to verify the number of constraints generated by the optimisation framework when the numbers of organs in each combinatory are increased. The chapter analyses the complementary and substitutable organs in the human system. The results of the constraint matrix generated from the CMAUT framework and the non-CMAUT data representation are compared. The Big O notation is used to verify the space complexity when clinical data is represented with the CMAUT framework as compared with non- CMAUT framework such as EAV and ERD techniques.

Again, in Chapter 6, the data size before optimisation and after optimisation were analysed using the CMAUT framework. The CVD clinical data of 400 randomly selected cohorts from the HSE, 2006 were used to conduct the statistical analysis. Two data sets were analysed statistically using the pair T-Test method to determine their P-value and Confidence Interval (CI) values. The resultant P-value was used to prove the hypothesis that “the difference between the data sizes before and after optimisation using the CMAUT framework is statistically significant or not. The graph of the data sizes before and after optimisation was drawn to further confirm the data reduction when the CMAUT framework is used.

6.1 CMAUT Frameworks

In CMAUT, the clinical data in the disease domain are modelled using UML class diagram and formalised using CMAUT linked with logical connectors as discussed in chapter 4 and 5. The output expression from the framework is written in mathematical format and optimised using LP based algorithm. The optimal results of the framework are used to determine the percentage risk of the participant getting the disease in the problem domain. Again, the output variables are mapped to the input multiple attributes of the clinical dataset of each participant. The framework is also used to analyse the clinical data and to determine the organ in the combination that the output has the optimal utility value in CMAUT or optimal attribute in *Non – CMAUT* system.

The CMAUT data representation technique in the framework is used to model clinical data and CMAUT algorithm is used to analyse the Big Data in clinical setting. The framework on receiving the input clinical data (i.e. attributes) perform the following operations:

1. Converts the various attributes into a common utility unit (U) using weights and multi-attribute utility theory (MAUT);
2. Attach the requisite organs in the logical statement to the U using the equivalent logical connector;
3. Then converts the logical expressions into conjunctive normal form (CNF);
4. Transforms the CNF into a set of inequalities or constraints matrix using the Table 6.1 in Raman, (1991);
5. Determine the organ that has the optimal value, the sum of the utility unit of each organ, becomes the coefficient of objective function of the organs in the Combinatory;
6. Determine the percentage risk of users; the framework has an inbuilt CMAUT algorithm that optimises the objective function using the LP technique. For the constraint matrix, all used attribute values are given 1 and the non-used attributes values are given 0.

The CMAUT framework was designed and developed using MATLAB 7.x optimisation toolbox. The difference between *non* – CMAUT and CMAUT are as follows: In this context, the CMAUT system is where the entities or objects in the disease domain are captured using class diagram and re-represented in CMAUT mathematical format. However, for the *non* – CMAUT system, an object or entity may have many attributes but these attributes are not converted into a common utility function or utility unit as proposed in this research. Some examples of the Non-CMAUT data representations are FOL/ERD and EAV/CR, which were discussed in chapter 4.

6.1.1 Space Complexity and the Application of Mathematical Operation

Mathematical operation was used to determine the space complexity of the new CMAUT data re-representation techniques for the complementary and substitutable organs. The big O function and notation were used to evaluate the space complexity of the CMAUT and non-CMAUT data representation.

To determine the space complexity the following mathematical operation was performed: The attributes in the algebraic expression were converted into utility unit as $X_1: [(C_1 \wedge C_2), U_1, U_2]$. The expressions were transformed into conjunctive normal form (CNF) and a set of inequalities (aka constraint matrix) using equation calculus and Raman's transformation Table 6.1 Raman et al, (1991) instead of $[0, 1]$ matrix used. The optimisation is carried out by maximising the objective function $X = \sum_i^n U_i C_i$ subject to the generated constraint matrix. The output of the algorithm and the constraint matrix are optimised using LP technique to determine the MAX clinical data required for CVD clinical investigation.

For ease of computation, the clinical object and components in this chapter are represented with Q_i for the mathematical operation. First the mathematical operation is conducted for two complementary organs and the numbers of constraints generated through the mathematical operation are entered into Table 6.2 below. Then the numbers of complementary organs were increased from 3 to 17 and the numbers of inequalities (aka constraints) generated are entered into the Table 6.2. Finally, the MATLAB software was used to plot a graph of the numbers of organs against numbers of constraints generated see Figure 6.1. In order to determine the space complexity of the complementary organs, the equation of the graph was generated using the BEST FIT function in MATLAB 7.x.

According to Fenton et al. (1997) the space complexity of the operation is determined by the highest power value specified in the BEST FIT equation. This is written as $Y = O(X^n)$ where n is the highest power in the equation and the concept is known as the big O notation. This method is used in the mathematical operations discussed in the section 6.2 below to determine the space complexity of the complementary and substitutable organs. This method is used to analyse the CMAUT data representation and those that does not use CMAUT approach are called Non-CMAUT in this research.

6.2 The Mathematical Operation Procedures

In this section is a summary of the mathematical algorithm and operation conducted and the reasons for carrying out the procedure. The steps used for the mathematical operation are:

1. Determine the number of constraints generated by the algorithm as the number of organs in CMAUT system increases and compare the results with those from *Non – CMAUT* (this denotes comparing the space complexity of CMAUT and *Non – CMAUT*);
2. Confirm the constraints (i.e. inequalities) obtained using CMAUT data re-representation mechanism and optimisation algorithm in the framework in section 6.2.1. This is used to determine the organs with optimal value in the CMAUT and *Non – CMAUT* system;
3. Determine the attribute values in the combinatory that needs further clinical analysis and use the CMAUT framework to find the predictive percentage risk in CVD CIS ;
4. Conduct benchmark analysis, to determine the consistency of the resultant percentage risk obtained from the CMAUT framework;

6.2.1 Conversion of the CMAUT logical expressions into Set of Inequalities:

The basics of the data re-representation mechanism in the CMAUT framework are as follows: first convert the different attributes into a common attribute utility unit (U_i). Then use the logical equivalence operator to rewrite the combinatorial logical expressions such that each organ in the combinatory has an equivalent utility unit attached to it that add up to the overall utility unit. For example, the CMAUT expression $[(Q_1 \text{ AND } Q_2), P_1, V_1, P_2, V_2]$ is rewritten using the equivalence operator as: $[Q_1 \ Q_2 \Leftrightarrow U_3]$. Inferring from section 6.1, it is seen that U_3 in the expression is the overall utility unit of all the attributes in the combinatory. Thus if U_1 is the utility unit for organ Q_1 and U_2 is utility unit for organs Q_2 and then the overall utility unit for the combinatory is $U_3 = U_1 + U_2$.

6.2.2 Algorithm for Conversion of CMAUT expression to Set of Inequalities

First the multiple attributes in CMAUT expression is converted into utility units using the weight allocated to each attribute and the following equation (6.1).

$$U = \text{sum} [w_i f(s)] \quad \text{where } f(s) = \frac{P_i - P_o}{P_o} \quad (6.1)$$

In this expression P_o is the standard recommended value and P_i is the actual measured value. An example of the conversion process for CMAUT expression is as follows:

$$[(Q_1 \wedge Q_2 \wedge Q_3), P_1, V_1, T_1, P_2, V_2, T_2, P_3, V_3, T_3]$$

Using the principle and formula (6.1) for converting the attributes to utility units, gives:

$$U_1 = \sum w_p f(sp_1) + w_v f(sv_1) + w_t f(st_1)$$

$$U_2 = \sum w_p f(sp_2) + w_v f(sv_2) + w_t f(st_2)$$

$$U_3 = \sum w_p f(sp_3) + w_v f(sv_3) + w_t f(st_3)$$

The w_p is the weight allocated to blood pressure, w_v , w_t are weights for blood volume and Total cholesterol. For mathematical operation reasons, the arithmetical sum of all the individual weights placed on all the attributes in the expression should be equal to 100.

$$[(Q_1 \wedge Q_2 \vee Q_3), U_1, U_2, U_3]$$

$$U = U_1 + U_2 + U_3$$

$$[(Q_1 \wedge Q_2 \vee Q_3), U]$$

$$[(Q_1 \wedge Q_2 \vee Q_3) \Leftrightarrow U]$$

The CMAUT expression in the framework is transformed into mathematical mixed integer program (MIP) for analysis. In this chapter 6, an algorithm is developed to convert the expression into conjunctive normal form (CNF), which are translated into set of inequalities for analysing the expressions.

The proposed algorithm converts the logical CMAUT expressions into a set of inequalities, using the Table 6.1 from Raman et al, (1991) to facilitate the transformation. These inequalities are used for evaluating the amount of memory space the data occupies.

The algorithm is:

1. Eliminate all the equivalences (\Leftrightarrow) and implications (\Rightarrow) by using the correspondence expressions in terms of AND, OR, and NOT;
2. Move the \rightarrow connective inside the brackets (apply the De Morgan's Law) For example $\rightarrow (Q_1 \vee Q_2) = \rightarrow Q_1 \vee Q_2$
3. Distribute the OR (\vee) over AND (\wedge). This is done by applying the distribution law i.e. $(A \wedge B) \vee C = (A \vee C) \wedge (B \vee C)$. For example $Q_3 \vee (Q_1 \wedge Q_2) = (Q_3 \vee Q_1) \wedge (Q_3 \vee Q_2)$.

4. Repeat steps 1, 2 and 3 until the literals are connected by \vee (OR) in the brackets and between the brackets are \wedge (AND).
5. Use the Table 6. 1 below to convert each of the expressions in the brackets or (the conjunctive clause) to inequalities.

Table 6.1 Representation of logical relations with linear inequalities Source (Raman et al., 1991):

Logical relation	Pure logical expression	Representation as linear inequalities
Logical “(OR)”	$Q_1 \vee Q_2, \dots, Q_n$	$q_1 + q_2 + \dots + q_n \geq 1$
Logical “(AND)”	$Q_1 \wedge Q_2, \dots, Q_n$	$q_1 \geq 1; q_2 \geq 1; q_n \geq 1$
Implication \Rightarrow	$Q_1 \Rightarrow Q_2$	$1 - q_1 + q_2 \geq 1$
Equivalence \Leftrightarrow	$(Q_1 \Leftrightarrow Q_2)$	$q_1 - q_2 \leq 0; q_2 - q_1 \leq 0;$ $q_1 = q_2$
Exclusive OR/ XOR	$Q_1 \oplus Q_2, \dots, Q_n$	$q_1 + q_2 + \dots + q_n = 1$

6.3 Generation of Constraints for Complementary organs using CMAUT

In this mathematical operation the complementary CMAUT expression is used to determine the number of constraints that $[(Q_1 \text{ AND } Q_2) \text{ equivalent } U_3]$, will generate when the CMAUT transformation algorithm is applied. The number of constraints generated is recorded in Table 6.2. The mathematical operation is repeated and the number of organs in the expression changed and likewise the results are recorded in Table 6.2 below.

6.3.1 Generation of Constraints for Complementary organs with CMAUT

- CMAUT Complementary system with two organs:

Example 1: The CMAUT expression for the complementary organs is $[(Q_1 \text{ AND } Q_2), P_1, V_1, P_2, V_2]$. The attributes P_1, V_1, P_2, V_2 are converted into attribute utility units (U_1 and U_2) using the utility formula 6.1. The expression is written as $[(Q_1 \text{ AND } Q_2) \text{ equivalent } U_3]$ and converted from $[(Q_1 \text{ AND } Q_2) \text{ equiv } U_3]$ into CNF and using Table 6.1. The transformation operation from CNF into set of inequalities (or constraints) is as follows:

$$[(Q_1 \text{ AND } Q_2), U_1, U_2] \text{ and then express as } [(Q_1 \text{ AND } Q_2), \Leftrightarrow U_3].$$

The CMAUT expression is transformed into logical statement using equivalence operator and proposition with pure logic operators;

$$\begin{aligned} & ((Q_1 \vee Q_2) \Rightarrow U_3) \wedge (U_3 \Rightarrow (Q_1 \wedge Q_2)) \\ & (\neg (Q_1 \vee Q_2) \vee U_3) \wedge (\neg U_3 \vee (Q_1 \wedge Q_2)) \end{aligned}$$

Then the DeMorgan's law is use to remove the bracket, so that it gives CNF as below:

$$(\neg Q_1 \vee \neg Q_2 \vee U_3) \wedge (\neg U_3 \vee Q_1) \wedge (\neg U_3 \vee Q_2) \quad (6.2)$$

To transform the clauses in the CNF expression into set of inequalities, first each expression in the clause is compared with those in Table 6. 1. Then the pure logic operators with literals are converted into mathematical expression with variables, by using the Table 6.1. This gives the set of inequalities below.

The first clause in the expression (6.2) is converted into a set of inequalities, using the following operation:

$$\neg Q_1 \vee \neg Q_2 \vee U_3$$

$\neg q_1 - q_2 + u_3 \geq 1$, multiplying $-ve$ on both sides of the expression and change the signs, the following is obtained:

$$q_1 + q_2 - u_3 \leq 1 \quad (6.3)$$

The second clause in the expression (6.2) is converted into a set of inequalities as follows:

$$\neg U_3 \vee Q_1$$

$$1 - u_3 + q_1 \geq 1 ; \quad -u_3 + q_1 \geq 0$$

$$q_1 - u_3 \geq 0 \quad (6.4)$$

The third clause in the expression (6.2) is converted into a set of inequalities as follows:

$$\neg U_3 \vee Q_2$$

$$1 - u_3 + q_2 \geq 1 ; \quad -u_3 + q_2 \geq 0$$

$$q_2 - u_3 \geq 0 \quad (6.5)$$

When all the set of inequalities from (6.3), (6.4), (6.5) are put together the following inequalities are obtained.

$$q_1 + q_2 - u_3 \leq 1$$

$$q_1 - u_3 \geq 0$$

$$q_2 - u_3 \geq 0$$

- CMAUT Complementary system with Three (3) organs

Example 2: The same procedure was used to convert the CMAUT expression with three organs, which is $[(Q_1 \text{ AND } Q_2 \text{ AND } Q_3), P_1, V_1, P_2, V_2, P_3, V_3]$ into a set of inequalities. First convert the attributes $P_1, V_1, P_2, V_2, P_3, V_3$ into utility units (U_1 and U_2 and U_3) using the utility formula (6.1). The resultant expression $[(Q_1 \text{ AND } Q_2 \text{ AND } Q_3) \text{ equiv } U_4]$ is transformed into CNF and then into a set of inequalities (or constraints) using Table 6.1. The result of simplifying the clauses using the algorithm in section 6.2 is as follows:

$$\begin{aligned} q_1 - u_1 &\geq 0 \\ q_2 - u_1 &\geq 0 \\ q_3 - u_1 &\geq 0 \\ q_1 + q_2 + q_3 - u_1 &\leq 2 \end{aligned}$$

In Appendix 6 are examples of how the algorithm in section 6.2 and Table 6.1 were used to convert the CMAUT expressions into set of inequalities with explanation.

- Results from generating the constraints for CMAUT complementary organs

Using the expressions in examples 1 and 2, series of mathematical operations were conducted to determine the numbers of inequalities generated for each set of complementary organs and their attributes. The numbers of organs were changed in the expression and the numbers of inequalities generated were recorded in Table 6.2. The results in Table 6.2 were used to draw the graph in Figure 6.1.

Table 6.2: CMAUT complementary organs

No. of complementary (AND) organs	No. of laterals including the Utility unit	No. of constraints (inequalities)
2	3	3
3	4	4
4	5	5
5	6	6
6	7	7
7	8	8
8	9	9
11	12	12
12	13	13
15	16	16
16	17	17
17	18	18

It is observed from the computed data and the graph drawn that the relationship between the numbers of organs in the combinatory and inequalities generated is $X \rightarrow x + 1$, $y = 1x + 1$. This is shown in Figure 6.1, where the results of constraints CMAUT CIS generated are plotted. The results are discussed in chapter 9.

$$X \rightarrow 1 * x + 1, y = 1 * x + 1$$

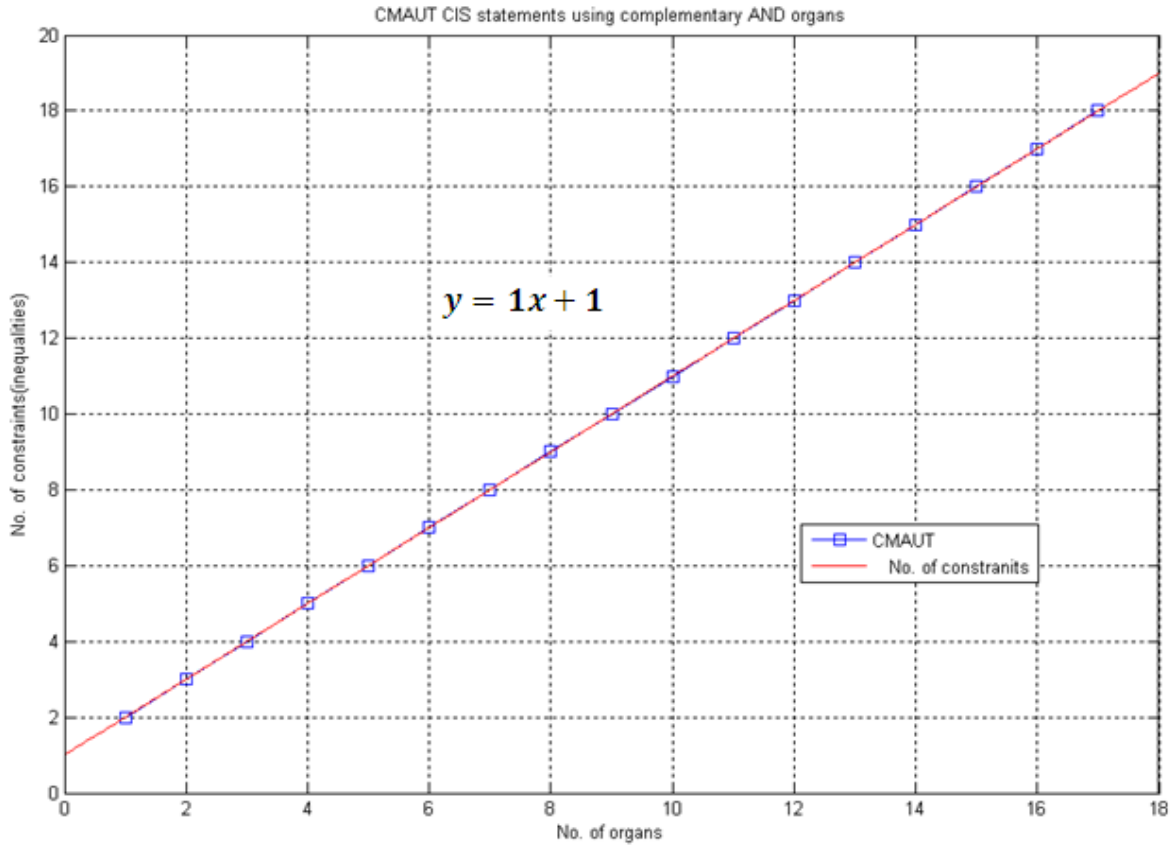


Figure 6.1: Graph of number of organs against number of constraints generated;

6.3.2 Generation of Constraints for Complementary organs with Non- CMAUT

In the *Non – CMAUT*, the CMAUT expression has only one attribute, which may be either blood pressure or total cholesterol as explained in section 6.3. In this research, only blood pressure values are used and therefore in the mathematical operation and algorithm every expression has pressure P_1 attached to the organ.

- Non- CMAUT Complementary system with two organs

Example 3: In this example, there are two organs with individual attribute that is pressure P, therefore the expression would be $[(Q_1 \text{ equiv } P_1) \text{ AND } (Q_2 \text{ equiv } P_2)]$ and the result will be as follows:

1. $(Q_1 \text{ OR NOT } P_1)$
2. $(P_1 \text{ OR NOT } Q_1)$
3. $(Q_2 \text{ OR NOT } P_2)$
4. $(P_2 \text{ OR NOT } Q_2)$

Transforming the clauses in the CNF above into a set of inequalities, using Table 6.1, the results will be:

$$\begin{aligned} q_1 - p_1 &\geq 0 \\ p_1 - q_1 &\geq 0 \\ q_2 - p_2 &\geq 0 \\ p_2 - q_2 &\geq 0. \end{aligned}$$

- Non- CMAUT Complementary system with three organs

Example 4: For the purpose of comparison the attribute used in this example is the blood pressure (P), which is set to be equivalent to the utility unit of the organ. This denotes that the utility unit U_i is computed with the blood pressure and the expression used is:

$[(Q_1 \text{ equiv } P_1) \text{ AND } (Q_2 \text{ equiv } P_2) \text{ AND } (Q_3 \text{ equiv } P_3)]$ and instead of the attribute P_i the U_i is used and the result of eliminating the high level operations shown below as:

$$\begin{aligned} &(((Q_1 \text{ AND } U_1) \text{ OR } (\text{NOT } Q_1 \text{ AND } \text{NOT } U_1)) \text{ AND } ((Q_2 \text{ AND } U_2) \text{ OR } \\ &(\text{NOT } Q_2 \text{ AND } \text{NOT } U_2)) \text{ AND } ((Q_3 \text{ AND } U_3) \text{ OR } (\text{NOT } Q_3 \text{ AND } \text{NOT } U_3))) \end{aligned}$$

To transform the clauses in the CNF into a set of inequalities, Table 6.1 was used and the results are presented below as:

$$\begin{aligned} q_1 - u_1 &\geq 0 \\ u_1 - q_1 &\geq 0 \\ q_2 - u_2 &\geq 0 \\ u_2 - q_2 &\geq 0 \\ q_3 - u_3 &\geq 0 \\ u_3 - q_3 &\geq 0 \end{aligned}$$

It is observed that in all the transformation operation the numbers of clauses in the CNF expression are equal to the number of inequalities the CMAUT expressions generate. In Appendix 6 are examples of how the algorithm in section 6.2 and Table 6.1 were used to convert the CMAUT expressions into set of inequalities with explanation.

- Results of Constraints generated for Complementary organs in *Non* – CMAUT

Using the expressions in Examples 3 and 4, series of mathematical operations were conducted to determine the numbers of inequalities generated for each set of complementary organs and their attributes using Non-CMAUT system. Then the numbers of organs were change in the expression and the numbers of inequalities generated are recorded in Table 6.3. The computed results in Table 6.3 were used to draw the graph in Figure 6.2.

Table 6.3: *Non* – CMAUT based CIS using complementary organs:

No of complementary (AND) organs	No of laterals including the Utility unit	No of constraints (inequalities)
2	2	4
3	3	6
4	4	8
5	5	10
6	6	12
7	7	14
8	8	16

It is observed from the computed data and the graph plotted that the relationship between the numbers of organs and the inequalities generated is: $X \rightarrow 2x + 1.4^{-15}$, $y = 2x + 1.4^{-15}$. This is shown in the graph Figure 6.2, where the results of Non-CMAUT CIS are plotted. It is subsumed from this equation that the space complexity for non-CMAUT is $X \rightarrow 2x$ or $y = 2x$ or $y = 2 * x$.

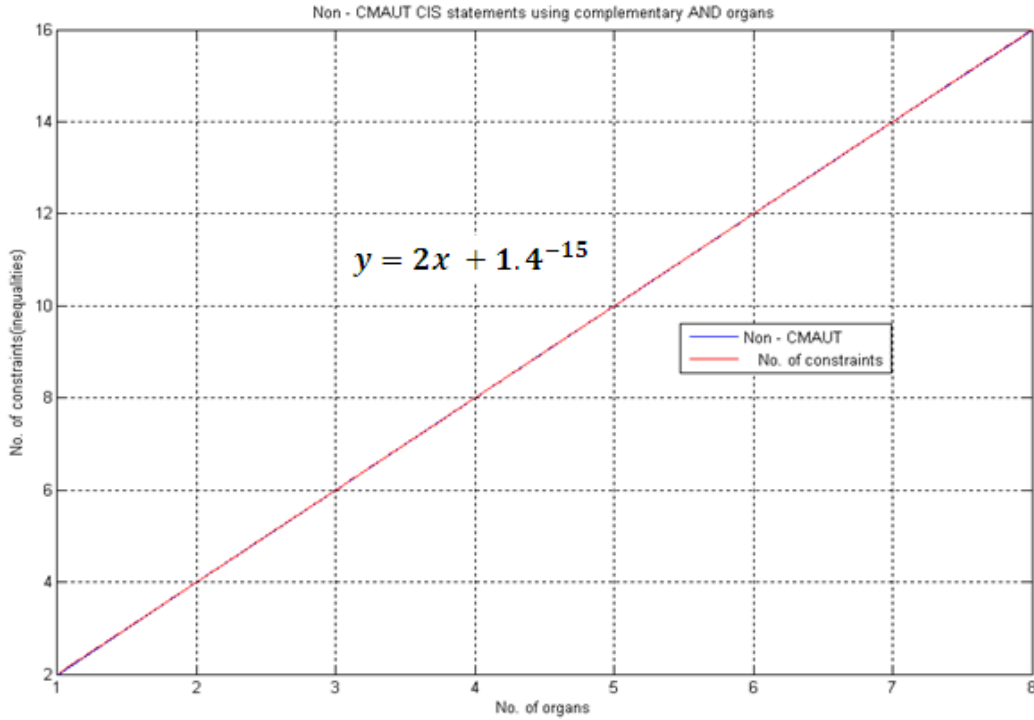


Figure: 6.2: Shows a comparison of *Non* – CMAUT using complementary organs (*AND*).

The comparison between Non-CMAUT and CMAUT data representation indicates that the constraints generated as the numbers of organs are increased are higher in Non-CMAUT than in the CMAUT system. The computed results are used to plot the comparative graph in Figure 6.2, which is evaluated in chapter 9.

6.4 Generation of Constraints for substitutable organs with CMAUT

Unlike the AND systems, the substitutable organ comprises of two or more sub organs that work in place of the other when one is malfunctioning, which subsumes one organ can stand in for the other. Therefore the individual attributes and utility units are unique to each organ. Hence they are not the arithmetical sum of the utility units of each organ. An example is the two kidneys in the human body they work independently to regular the flow of fluid in the human being (Guyton et al. 2006). In CMAUT representation they are modelled as subclasses in Figure 4.7. The CMAUT expression for substitutable organ is written as follows:

$$[(Q_1 \text{ OR } Q_2), P_1, V_1, P_2, V_2]$$

Using the utility formula 6.1 the expression can be rewritten as $[(Q_1 \vee Q_2), \Leftrightarrow U_3]$, which represents one complete combinatorial clinical data. In this expression, the flow of attributes from the two organs must meet at one joint so the overall utility unit will be U_3 and under this condition, the attributes are measured at the junction. The second method of expressing the

substitutable in CIS is to attach the individual utility unit to their respective organs. This is written as $[(Q_1 \Leftrightarrow U_1) \text{ OR } (Q_2 \Leftrightarrow U_2)]$ or in other words as $[(Q_1 \Leftrightarrow U_1) \vee (Q_2 \Leftrightarrow U_2)]$. This expression denotes that each independent organ has their utility unit. For example, organ1 has a utility value of U_1 and organ 2 has utility value of U_2 , in this scenario their total utility unit cannot be the sum of the utility values of the individual organs.

6.4.1 Substitutable organs using CMAUT Framework

This section explains the two ways of expressing the substitutable organs in CMAUT. The first operation uses the expression $[(Q_1 \text{ OR } Q_2)P_1, V_1, P_2, V_2]$ to generate the constraints. This is followed by the second option that uses the expression:

$(Q_2 \text{ equiv } U_1) \text{ OR } (Q_2 \text{ equiv } U_2) \text{ OR } (Q_3 \text{ equiv } U_3)$ for both CMAUT and non-CMAUT.

The second expression is used for both CMAUT and non-CMAUT because from the discussion in chapter 4, it is subsumed that both systems use the same data representation.

The first method of expressing substitutable organs in CMAUT using *OR* logical expression is presented below as:

$$[(Q_1 \text{ OR } Q_2)P_1, V_1, P_2, V_2]$$

$$[Q_1 \vee Q_2 \Leftrightarrow U_3]$$

Transforming the equivalence operators to pure logic operators would give the following:

$$(Q_1 \vee Q_2 \Rightarrow U_3) \wedge (U_3 \Rightarrow Q_1 \vee Q_2)$$

$$(\neg (Q_1 \vee Q_2) \vee U_3) \wedge (\neg U_3 \vee (Q_1 \vee Q_2))$$

The DeMorgan law is used to remove the bracket, so that it gives the CNF below:

$$((\neg Q_1 \wedge Q_2) \vee U_3) \wedge (\neg U_3 \vee Q_1 \vee Q_2)$$

$$(\neg Q_1 \vee U_3) \wedge (\neg Q_2 \vee U_3) \wedge (\neg U_3 \vee Q_1 \vee Q_2)$$

The set of inequalities with integer variables that is computed would be as follows:

$$q_1 + q_2 - u_3 \geq 0$$

$$u_3 - q_1 \geq 0$$

$$u_3 - q_2 \geq 0$$

- CMAUT substitutable system with two organs using the second approach:

This research uses the second approach for analysing the substitutable organs where the individual utility units are attached to their respective organs. This expression is written as $[(Q_1 \Leftrightarrow U_1) \text{ OR } (Q_2 \Leftrightarrow U_2)]$ or in other words as $[(Q_1 \Leftrightarrow U_1) \vee (Q_2 \Leftrightarrow U_2)]$.

Example 5: This second method is used for both CMAUT and *non* – CMAUT systems where the individual organs have their own blood pressure values or utility units attached to them. This example 5 is based on CMAUT data representation with 2 organs and it is written as: $[(Q_1 \text{equiv } P_1, V_1) \text{ OR } (Q_2 \text{equiv } P_2, V_2)]$. The attributes in the expression are converted into utility units, which are U_1 and U_2 .

The final expression is written as: $[(Q_1 \text{equiv } U_1) \text{ OR } (Q_2 \text{equiv } U_2)]$, which is expressed as:

$$[(Q_1 \Leftrightarrow U_1) \text{ OR } (Q_2 \Leftrightarrow U_2)] \text{ or } [(Q_1 \Leftrightarrow U_1) \vee (Q_2 \Leftrightarrow U_2)].$$

When this expression is simplified the numbers of clauses obtained are:

$$((Q_1 \text{ OR } Q_2 \text{ OR NOT } U_1 \text{ OR NOT } U_2) \text{ AND } (Q_1 \text{ OR } U_2 \text{ OR NOT } U_1) \text{ AND } (Q_2 \text{ OR } U_1 \text{ OR NOT } Q_1 \text{ OR NOT } U_2) \text{ AND } (U_1 \text{ OR } U_2 \text{ OR NOT } Q_1 \text{ OR NOT } Q_2))$$

Therefore the list of clauses will be as follows:

$$\begin{aligned} &(Q_1 \text{ OR } Q_2 \text{ OR NOT } U_1 \text{ OR NOT } U_2) \\ &(Q_1 \text{ OR } U_2 \text{ OR NOT } Q_2 \text{ OR NOT } U_1) \\ &(Q_2 \text{ OR } U_1 \text{ OR NOT } Q_1 \text{ OR NOT } U_2) \\ &(U_1 \text{ OR } U_2 \text{ OR NOT } Q_1 \text{ OR NOT } Q_2) \end{aligned}$$

The set of inequalities with integer variables would be as follows:

$$q_1 + q_2 - u_1 - u_2 \geq -1$$

$$q_1 + u_2 - q_2 - u_1 \geq -1$$

$$q_2 + u_1 - q_1 - u_2 \geq -1$$

$$u_1 + u_2 - q_1 - q_2 \geq -1$$

- CMAUT substitutable system with three organs

Example 6: This second example converts three substitutable organs with different attributes, which are transformed to utility unit and then into a set of inequalities. The example uses the expression $[(Q_1 OR Q_2 OR Q_3)P_1, V_1, P_2, V_2, P_3, V_3]$ to generate the constraints. This expression is rewritten in CMAUT format as:

$$[(Q_1 equiv P_1, V_1) OR (Q_2 equiv P_2 V_2) OR (Q_3 equiv P_3 V_3)]$$

The attributes in the expression are converted into utility units, which are U_1 , U_2 and U_3 . In this research, the CMAUT expression for 3 organs is written using utility units as follows:

$$[(Q_1 equiv U_1) OR (Q_2 equiv U_2) OR (Q_3 equiv U_3)]$$

The set of inequalities with integer variables computed from the above expression would be

$$q_1 + q_2 + q_3 - u_1 - u_2 - u_3 \geq -2$$

$$q_1 + q_2 + u_3 - u_1 - q_3 - u_2 \geq -2$$

$$q_1 + q_3 + u_2 - q_2 - u_1 - u_3 \geq -2$$

$$q_1 + u_1 + u_3 - q_2 - u_1 - q_3 \geq -2$$

$$q_2 + u_1 + q_3 - q_1 - u_2 - u_3 \geq -2$$

$$q_2 + u_1 + u_3 - q_1 - q_3 - u_2 \geq -2$$

$$u_1 + q_3 + u_2 - q_1 - q_2 - u_3 \geq -2$$

$$u_1 + u_2 + u_3 - q_1 - q_2 - q_3 \geq -2$$

Examples of how the CMAUT algorithm in section 6.2 was used to convert the CMAUT expressions into set of inequalities are shown in Appendix 6.

- Results of constraints generated for substitutable organs (OR)

Using the expressions in Examples 5 and 6, series of mathematical operations were conducted to determine the number of inequalities that are generated for each set of substitutable organs and their attributes applying the CMAUT data representation. The numbers of organs in the expression are changed and the corresponding numbers of inequalities generated are recorded in Table 6.4. The results in Table 6.4, was used to draw the graph in Figure 6.4.

6.4.2 Constraints Generation for substitutable organs with Non- CMAUT

As explained in section 6.4.1, the same expression is used for both CMAUT and Non-CMAUT data representation. The expression used for Non-CMAUT is $(Q_1 \text{equiv } U_1) \text{OR } (Q_2 \text{equiv } U_2) \text{OR } (Q_3 \text{equiv } U_3)$, and the numbers of organs in the expression was increased at every stage during the computation exercises. Below in Table 6.6 are the results of the numbers of constraints generated and Figure 6.6 shows the graph that was plotted from the results. It was observed that for substitutable organs the relationship between the number of organs in the combinatory and the number constraints computed is $x \rightarrow 2x$ therefore $y = 2^x$. This is explained in the evaluation chapter 9.

Table 6.4: CMAUT using substitutable organs (OR) also called partial substitutable organs

No of substitutable (OR) organs	No of laterals including the Utility unit	No of constraints (inequalities)
2	2	4
3	3	8
4	4	16
5	5	32
6	6	64
7	7	128
8	8	256

- Non- CMAUT substitutable system with two organs (OR)

For Non-CMAUT CIS, the blood pressure values (P) were associated with each of the individual organs. This is unlike *CMAUT* where each substitutable organ has its own set of attributes, which are converted into utility unit. To prove that the results from the non-CMAUT and CMAUT are the same, the following example with the same expression is used.

Example 7:- A non-CMAUT with two organs and associated pressure is expressed as $[(Q_1 \text{equiv } P_1) \text{OR } (Q_2 \text{equiv } P_2)]$. The list of clauses after simplification is shown as:

$$\begin{aligned}
 &(Q_1 \text{OR } Q_2 \text{OR NOT } P_1 \text{OR NOT } P_2) \\
 &(Q_1 \text{OR } P_2 \text{OR NOT } G_2 \text{OR NOT } P_1) \\
 &(Q_2 \text{OR } P_1 \text{OR NOT } Q_1 \text{OR NOT } P_2) \\
 &(P_1 \text{OR } P_2 \text{OR NOT } Q_1 \text{OR NOT } Q_2)
 \end{aligned}$$

The set of inequalities with integer variables would be:

$$q_1 + q_2 - p_1 - p_2 \geq -1$$

$$q_1 + p_2 - q_1 - p_1 \geq -1$$

$$q_2 + p_1 - q_1 - p_2 \geq -1$$

$$p_1 + p_2 - q_1 - q_2 \geq -1$$

- Non- CMAUT substitutable system with three organs (OR)

Example 8:- Similarly the *Non – CMAUT* with three organs is expressed as $[(Q_1 \text{equiv } P_1) \text{OR } (Q_2 \text{equiv } P_2) \text{OR } (Q_3 \text{equiv } P_3)]$ and the number of clauses obtained after simplification is as follows:

The set of inequalities with integer variable expressions would be as follows:

$$q_1 + q_2 + q_3 - p_1 - p_2 - p_3 \geq -2$$

$$q_1 + q_2 + p_3 - p_1 - q_3 - p_2 \geq -2$$

$$q_1 + q_3 + p_2 - q_2 - p_1 - p_3 \geq -2$$

$$q_1 + p_1 + p_3 - q_2 - p_1 - q_3 \geq -2$$

$$q_2 + p_1 + q_3 - q_1 - p_2 - p_3 \geq -2$$

$$q_2 + p_1 + p_3 - q_1 - q_3 - p_2 \geq -2$$

$$p_1 + q_3 + p_2 - q_1 - q_2 - p_3 \geq -2$$

$$p_1 + p_2 + p_3 - q_1 - q_2 - q_3 \geq -2$$

Examples of how the algorithm in section 6.2 was used to convert the *Non – CMAUT* expressions into set of inequalities are shown in Appendix 6. Comparing the results from CMAUT with the results from *CMAUT* CIS, it is inferred that the clauses and the set of inequalities are the same. Therefore for substitutable organs, where OR is used, the same expression can be used for both *CMAUT* and *Non – CMAUT* data representation.

- Results of constraints generated for substitutable organs (OR) – Non CMAUT

Using the expressions in examples 7 and 8, series of mathematical operations were conducted to determine the numbers of inequalities generated for each set of substitutable organs and

their attributes in Non-CMAUT system. The numbers of organs in the expression were changed and the numbers of inequalities generated recorded in Table 6.5. The results in Table 6.5 were used to draw the graph in the Figure 6.4.

Table 6.5: Non-CMAUT using substitutable organs (OR) - partial substitutable organs

No of substitutable (OR) organs	No of laterals including the Utility unit	No of constraints (inequalities)
2	2	4
3	3	8
4	4	16
5	5	32
6	6	64
7	7	128
8	8	256

Figure 6.4; shows the comparison of the results obtained from using substitutable organs (OR) for Non- CMAUT and CMAUT data representations. From Figure 6:4, the substitutable organs have a relation of $x \rightarrow 2^x$ that is $y = 2^x$ this indicates that as the numbers of organs increase the numbers of constraints also increase. This is discussed in evaluation chapter 9.

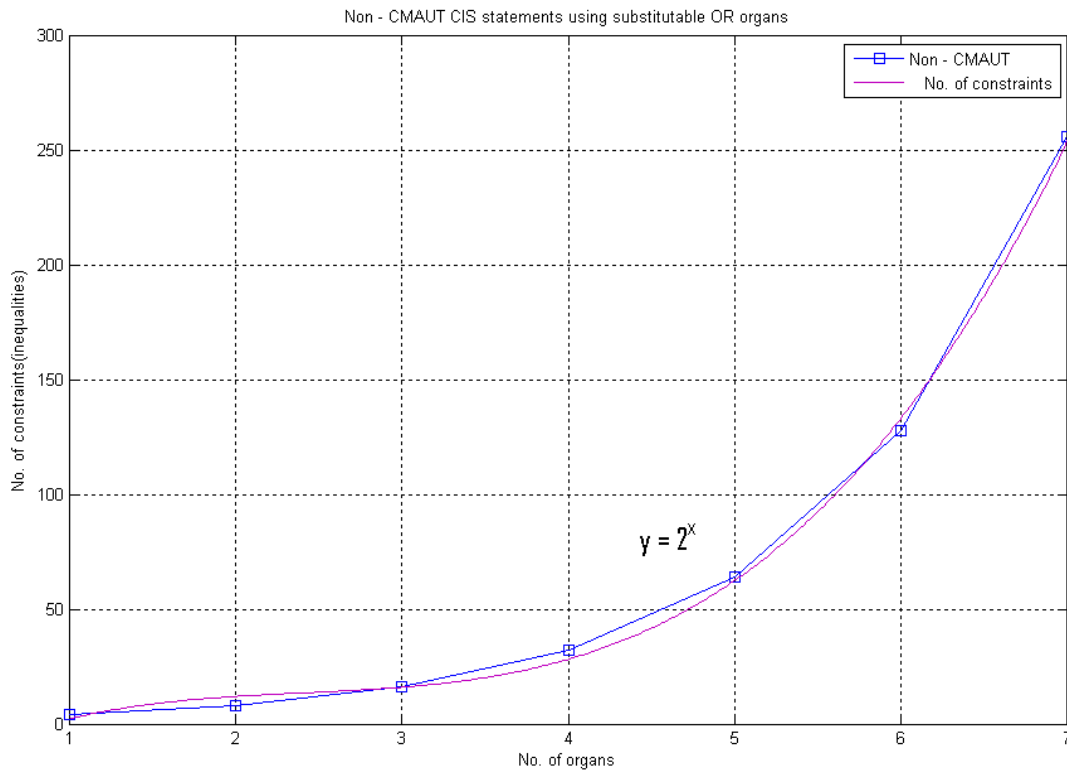


Figure 6.3: Non – CMAUT CIS statement using substitute OR organs

6.4.3 Summary of the Space Complexity for CMAUT Framework

From the graphs in Figure 6.1 and Figure 6.2, it is observed that for complementary organs when the numbers of organs in the combinatory are increased the number of constraints also increases. For complementary organs that use CMAUT data representation the computational space increases with function of $y = x + 1$, while CIS with Non-CMAUT data representation increases in order of $y = 2x$. Using the Big O notation, from the results of the functional performance test it is subsumed that the space complexity of CMAUT system is $\theta = (n + 1)$ as compared to the CIS system without CMAUT of $\theta(2n)$. This is discussed in Chapter 9.

Similarly, Figure 6.3 shows that when the numbers of organs and attributes in the substitutable combinatorial expression using the OR connector are increased the number of constraints also increases. Therefore for CIS where substitutable organs expression are used both in the CMAUT and Non CMAUT data representations have the same space complexity of $x \rightarrow 2^x$ which is $y = 2^x$. From the functional performance test results the space complexity of both substitutable organ data representation methods is $\theta(2^x)$. These results are discussed in the evaluation chapter 9.

6.5 Analysis of clinical data sizes before and after optimisation

For the analysis of the clinical data sizes before and after optimisation 402 participants were selected from the 3645 participants and their records retrieved from the (HSE, 2006) and stored in plain text format. This approach was used to avoid any increase of data size caused by the files format such as .doc, .docx etc. Each of the clinical data of the selected 402 participants was input into the CMAUT framework and executed to determine the APR values as well as their hypertension status. This is also used to identify the attributes in the combinatory that needs investigation.

The output file of each participant was stored in text file and the data size measured and recorded against the corresponding Participant Identification Number (PIND). The clinical data sizes before and after optimisation was input into an Excel file and presented in Table 6.6. This is followed by the determination of the P_value using the T-test in the SPSS package.

6.5.1 Statistical Analysis of data size before and after optimisation with CMAUT.

The hypothesis of this research states that; the application of CMAUT and logical connectors using mathematical expression that is optimised with LP technique will reduce the space complexity and amount of data required for CVD decision making. The first part of this chapter, explained how the use of CMAUT data re-representation and logical connectors improves the space complexity. This section focuses on the second part of the hypothesis that the optimisation framework reduces the amount of data needed to determine the percentage risk of a participant been hypertensive.

- Procedure for analysing data size reduction in CMAUT Framework:

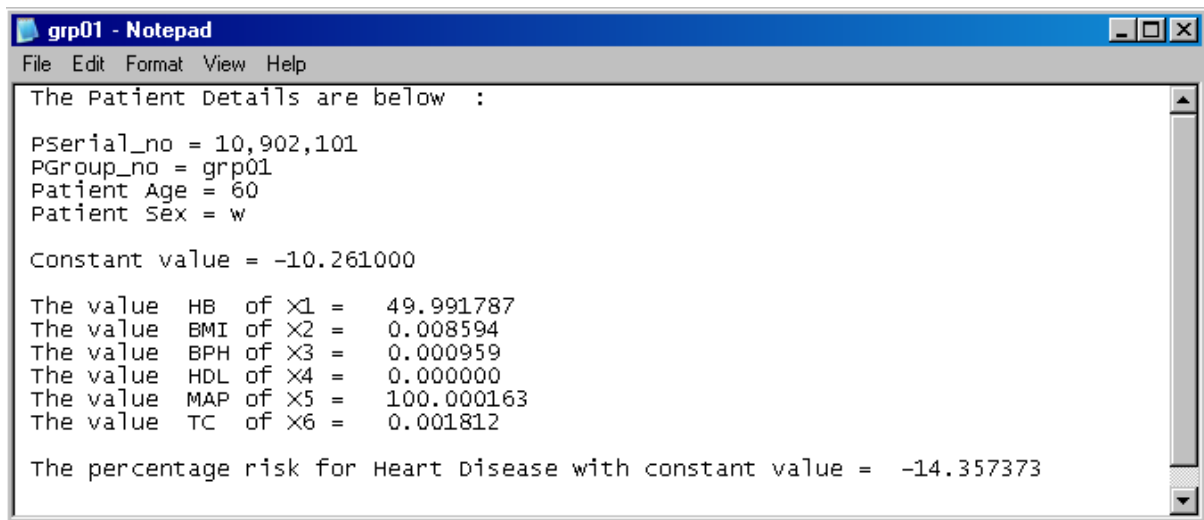
In chapter 5, it was established that the CMAUT model 2 will be used to analyse the CVD framework because it has better kappa and prevalence values as compared to the CMAUT model 1. This denotes that model 2 has a higher level of agreement with the GP diagnosis as compared with model 1. Based on this pretext, 10% of the 3645 participants, which is 364, were selected as the sample participants for the data size investigation and the data reduction trial. However for approximation purpose, the clinical data of 402 participants were used.

- Data size before optimisation

The CVD data for each of the 402 participants was exported from Microsoft Excel spreadsheet into text files. This is because text files eliminate all the formatting and configuration that are associated with the different types of files such as PDF, Excel, and Word. The data size of each participant before optimisation was input into text file and the results are measured and recorded. The size of each of participant's file before optimisation is recorded against the participant's serial number (Pserial no.) in Excel sheet as indicated in Table 6.6. The attribute values of the 402 participants before optimisation were entered into text file and measured on UNIX platform using the input command for the purpose of consistency. To compare the data sizes before and after optimisation the same measurement operation was performed after optimisation using the CMAUT framework and the values recorded. This is discussed below.

- Data size after optimisation

The CVD clinical record of each of the 402 participants was simulated using the CMAUT optimisation framework and the optimised output results stored in individual text files. The size of each participant's file after optimisation were stored in ZIP file that was exported into Excel sheet and recorded against the participant's Pserial number. The output records of all the 402 participants were input into text file after optimisation and measured. Figure 6.4 is one of patient's records, which is stored and displayed in text file format for data analysis.



```

File Edit Format View Help
The Patient Details are below :
PSerial_no = 10,902,101
PGroup_no = grp01
Patient Age = 60
Patient Sex = w
Constant value = -10.261000
The value HB of X1 = 49.991787
The value BMI of X2 = 0.008594
The value BPH of X3 = 0.000959
The value HDL of X4 = 0.000000
The value MAP of X5 = 100.000163
The value TC of X6 = 0.001812
The percentage risk for Heart Disease with constant value = -14.357373

```

Figure 6.4: Percentage risk for heart disease calculation on the text file.

6.5.2 Statistical Analysis of the Results using Pair T-test in SPSS

The pair T-test statistical method was used because this method allows the comparison of the two groups of data sizes, which maybe either continuous dependant or independent variables. Therefore, the difference between the data sizes of the selected 402 participants before and after optimisation using the CMAUT framework were verified with the pair T-test method. Again, the pair T-test technique was used instead of the ANOVA method because this research deals with two groups of data sizes of continuous dependant variables and not the comparison of three or more groups.

In this research, samples dependant pair t-test was used because the data sizes of the same participants were compared at different period (i.e. before and after optimisation).

This is different from independent sample t-test, which compares two different groups of participants (Campbell, 2006).

- Procedure for determining T-Test in SPSS

Figure 6.5 depicts the procedure used for conducting the pair sample T –test analysis. In SPSS version 15, the Analyse menu was selected and then the Compare Means clicked. This is followed by selecting the Pair samples T-Test in the drop down menu. There appears a screen in which the data sizes before and data sizes after optimisation appear as variables. The two variables were selected and moved into the Pair Variables box on the right side, and then the OK button was selected and clicked. The output Table 6.5 then appear as the resultant calculation from the IBM SPSS package.

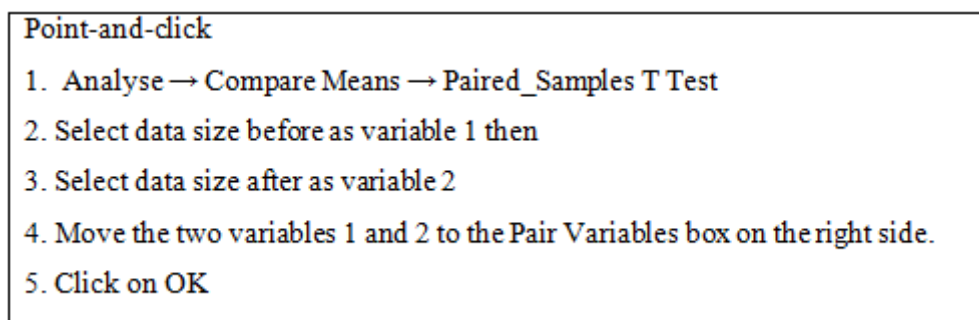


Figure 6.5: Procedure for conducting dependant samples t-Test in SPSS

- Output from SPSS and interpretation of the results

The results from the SPSS software after the operation are shown below Table 6.5. The SPSS software gave the statistical significant value of the P_value as 0.00 at a probability level (2-tailed). The significant P_value of 0.00 is less than the 0.01 or 0.001, which is the acceptable standard value in medical application (Campbell, 2006). Therefore the difference between file sizes before and after Optimisation is statistical significant.

Table 6.5 the Output results of the Paired Samples Statistics:

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	File size before Optimisation (In bytes)	1216.66	402	21.801	1.087
	File size after Optimisation (In bytes)	463.50	402	.916	.046

Paired Samples Correlations:

		N	Correlation	Sig.
Pair 1	File size before Optimisation (In bytes) & File size after Optimisation (In bytes)	402	-.029	.557

Paired Samples Test:

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Size before Optimisation size after Optimisation (In bytes)	753.162	21.847	1.090	751.020	755.304	691.219	401	.000

From Table 6.5, it was observed that the mean values of the data sizes before optimisation is 1216.66 bytes and after optimisation is 463.50 bytes. The standard deviation between the two sizes is 21.801 and 0.916 but there is a negative correlation between them. The data sizes after optimisation were found to be lower than the data sizes before optimisation with the calculated t score at 691.219, degree of freedom of 401. The statistical significant P value is less than 0.001 for the probability level of 2-tailed, which is expressed as ($t = 691.219$, $df = 401$, $p < 0.001$).

The CI for the clinical data before and after optimisation using the CMAUT framework is 751.020 for the lower boundary and 755.304 for the upper boundary. Therefore, the actual difference in value between the lower and upper boundaries is 753.16 Bytes. The T-test analysis gave a p-value of 0.000, which is less than 0.01, which denotes that the results are statistically significant. Hence the null hypothesis is rejected and the alternative hypothesis, which states that the clinical information can be optimised using CMAUT framework to reduce the amount of data required for primary care investigation thus reducing information overload is accepted. This is confirmed in Figures 6.6 and 6.7 and discussed in Chapter 9.

Table 6. 6A: Data size for 10 participants before and after optimisation with CMAUT

No. Of participants	Pserial no.	Data size before optimisation (bytes)	Data size after optimisation (bytes)
1	10,902,101.00	1256	465
2	10,846,103.00	1251	464
3	11,039,102.00	1251	463
4	11,046,101.00	1251	465
5	11,239,101.00	1245	464
6	11,249,102.00	1244	464
7	11,306,101.00	1249	464
8	11,313,101.00	1245	463
9	11,349,102.00	1243	464
10	11,356,101.00	1262	464

- Graphical Representation of the data size before and after optimization:

Using the data values from the Table 6.6B in the appendix and the information of all the 402 participants' data set in table 6.6C Appendix 6, the graphs in Figure 6.6 were drawn. From the individual graphs it is subsumed that the data sizes before optimisation that has a mean value of 1216.66 bytes are higher than the 463.50 bytes, which is the size after optimisation.

Again, the comparison graph in Figure 6.6 shows that there is a great difference between the two set of data, which is confirmed in the statistical analysis discussed during the T-Test. It is therefore subsumed that the difference between the data size before and after optimisation using the CMAUT is physical and statistically significant.

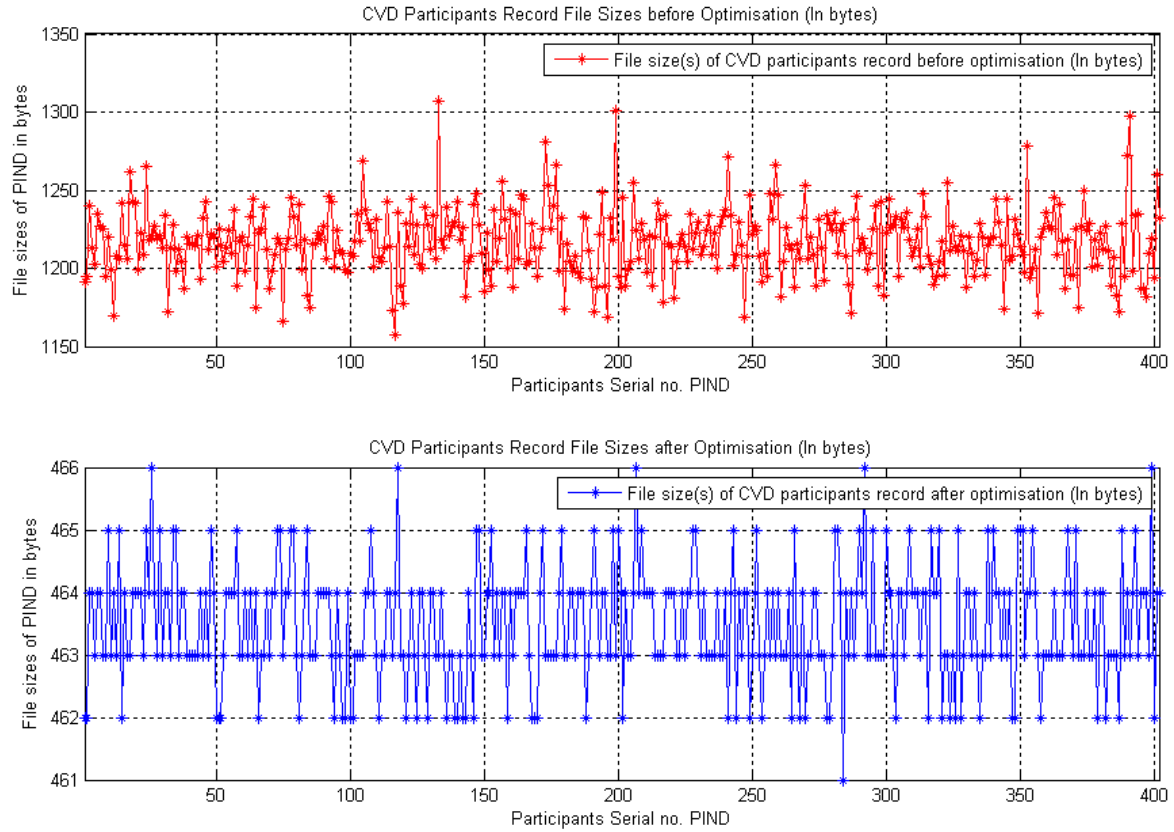


Figure 6.6: Before (red curve) and after optimisation (blue curve) of patient record data file

Figure 6.6 shows the results of the optimisation of 402 participants' records. Since the results were generated using the CMAUT model, the results are acceptable for the entire CVD participants' records. These results are further discussed in the evaluation chapter 9.

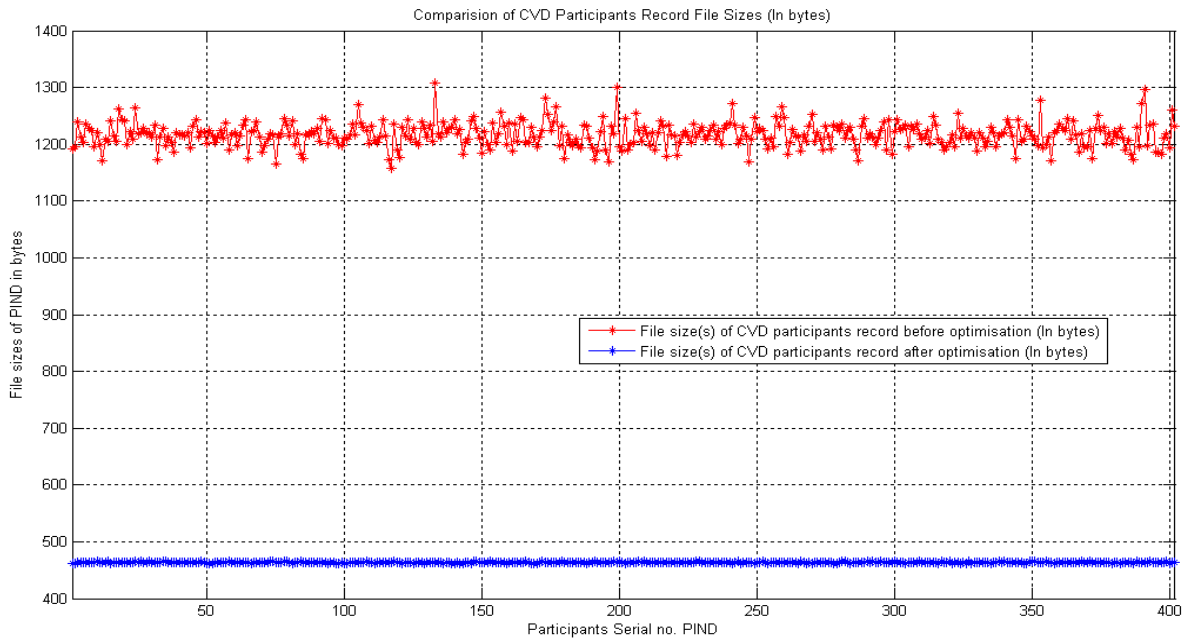


Figure 6.7: Comparison of CVD Participants records file sizes (in bytes)

6.6 Summary

The space complexity challenges of the CMAUT framework was analysed and confirmed in sections 6.3 and 6.4. It was established that the amount of clinical data required for CVD analysis and decision making using the CMAUT risk prediction model is significantly reduced in the CVD framework. This was proved using mathematical analysis, graphical representation and statistical significant techniques. From the space complexity analysis conducted, it was observed that the proposed CMAUT clinical data re-representation method generates less constraint expressions as compared to the non-CMAUT system for complementary of organs. However, for substitutable organs both systems generate equal amount of constraints. This is acceptable because in the human body there are more complementary organs compared with the substitutable organs (Guyton and Hall, 2006).

The statistical analysis revealed that using the CMAUT optimization framework reduces the data size that is retrieved from the data source for clinical decision making by approximately 700 Bytes for each participants. This is because the mean difference between the data sets before and after optimization is 753.16 Bytes. Therefore for a million patients' records that are transmitted every day this will lead to a reduction of approximately 700 million Bytes. This is a huge cost effective technique, which also reduces the information overload in the computer network and the amount of data that medics need for decision making. Finally this confirms the first part of the hypothesis, which states that the use of the CMAUT optimisation framework will reduce the clinical data size and space complexity required for decision making in CIS. This is further discussed in the evaluation chapter 9.

Chapter 7: CMAUT CVD Risk Prognosis Framework

7.0 Introduction:

This chapter discusses the second part of the hypothesis, which states that CMAUT CVD optimisation framework, can be used to predict the percentage risk (PPR) of a user been hypertensive. The modelling of the Prognosis framework and the simulations are carried out using the CVD data collected from the HSE survey conducted in 2006 (HSE, 2006). This chapter is related to chapter 5, where the diagnosis framework was discussed. In this chapter two prognosis models are designed and implemented in MATLAB and their outputs are verified using the same 3654 participants from HSE survey data used in Chapter 5. The PPR results in chapter 7 are compared with the results from chapter 8 and evaluated in chapter 9.

7.1 CVD Predictive Percentage Risk and CMAUT Prognosis framework

According to Jackson (2005), the predictive risk of cardiovascular disease is the probability that an individual will have cardiovascular event in a specific period. In this definition, an event denotes the chances that a participant will have cardiovascular disease in 5 or 10 years depending on their CVD risk factors and predictors condition. In prognosis, PPR is defined as the measure of the likelihood of a patient developing a disease over a time period, while relative risk measures the chances of risk occurring in two different groups of people.

As discussed in Chapter 5, there are two types of CVD predictive risk namely the clinical initial absolute percentage risk (APR), which is used to estimate the likelihood of user been hypertensive based on their current measurable clinical parameters. This evidence based clinical prediction model is used for clinical diagnosis and it is based on measurable clinical parameters. The measurable parameters used in this research are blood pressure (HB), BMI, BPH, BPL, MAP, HDL and TC. The diagnosis framework in Chapter 5 uses only measureable attribute values for the computation of the absolute percentage risk (APR) (Panagiotakos and Stavrinos, 2006).

The second type of CVD risk predictor is the predictive percentage risk (PPR), which is the measure of the chances that an individual will develop CVD disease over a specified time based on their measurable and non-measurable attributes or risk factors (Campbell, 2006).

The PPR values are used in prognosis to forecast the chances of individuals or group of people in a population becoming hypertensive over a period of time. This is also used as an epidemiological tool for CVD risk prediction (Sanderson, 2007).

In this research, for the purpose of designing the CVD prognosis framework two types of CVD risk predictors namely the Web based heart risk calculators and Framingham Risk equations were analysed (Wilson P, 1998) (Brindle 2003). It was identified that these two types of prediction models use both measureable and non-measureable attributes. The risk predictors used in these prognosis models are Age, Sex, HB, BMI, BPH, BPL, HDL, MAP, Diabetic, Total Cholesterol, Smoking, Existing CVD and ECG (Sheridan et al, 2003). Therefore this research uses the same risk predictors for verification proposes. For consistency the predictive percentage risk in 10 years is the arithmetic sum of the absolute percentage risk (APR) and the computed predicted risk in 10 years (Edoh et al. 2011).

- CMAUT Diagnosis framework and Absolute Percentage Risk (APR)

In chapter 5, the operation of the CMAUT framework was discussed and it is summarised as follows: - The framework comprises of two subsystems, which are the CMAUT clinical Data Re-representation mechanism that uses class model and LP based CMAUT optimisation algorithm implemented in MATLAB. The data re-representation mechanism is made up of subsystems that capture the clinical data using class model. The CMAUT system re-represents the clinical data in logical format and formalise it for mathematical manipulation and analysis. In this subsystem, only the measurable clinical parameters discussed above are modelled and used to determine the beta coefficients. These coefficient values constituent the weighting applied in the utility function and in the formulation of the objective function.

The optimisation algorithm: - The CMAUT logical expressions from the data re-representation mechanism are converted into mathematical formalisation, which serves as input into the algorithm. This research focuses on complementary organs, where the utility unit of the individual attributes are arithmetically sum together to create the objective function to be optimised. This is written as $Z = \sum_i^n (U_1X_1 + U_2X_2 + \dots + U_nX_n)$. This objective function is optimised subject to the generated unit matrix. In this algorithm the attributes that are used in the inequality matrix are depicted as 1 and those not as 0.

The optimisation algorithm was written in MATLAB and used to determine the clinical absolute percentage risk (APR) of a user been hypertensive. In prognosis, the initial absolute percentage risk (APR) is known as the Initial clinical percentage risk value, which is denoted as (u). The framework allows each output variable to be mapped to their corresponding attribute, which enables the decision maker to identify the attributes in the combinatory that needs to be analysed for further medical investigation. The CMAUT framework and its operation were discussed in chapter 5.

7.2 The CMAUT Prognosis Framework

The CIS prognosis framework is made up of the data re-representation mechanism and the LP optimization algorithm, which is designed using the utility function. In this framework, the probability of a disease occurring at a specified time is identified by the abnormal values of measurable and non-measurable attributes from the specified norm (i.e. baseline) in the disease domain. The method used to model and re-represent data is the same as discussed in Chapter 5 under the CMAUT diagnosis framework. The relationship between the organs in the disease domain is described using combinatorial logical connectors and the attributes of the organs expressed with multiple attributes as in UML class model. The multiple attributes in the expressions are used to calculate the utility unit (U) of each of the organ using the Utility Unit formula. The CMAUT expression formulated from the multiple attributes in the class diagram are converted into mathematical format, which serves as input to the LP optimisation algorithm. The output percentage risk value of the first part of the framework is the Initial clinical percentage risk value which is (u).

In prognosis framework, both the measurable and non-measurable CVD risk factors are used to calculate the predictive time value $P(T)$. The resultant percentage risk is the added of the Initial clinical percentage risk value (u) and the predictive risk factor $P(t)$ to obtain the predictive percentage risk $P(T)$ in 10 years for each participant.

The stepwise procedure for computation of CVD Predictive Percentage Risk is written in the Box below:

1. Capture the relations to be optimized in the disease domain with UML- CMAUT
2. Identify the measurable Risk factors or attributes in the problem domain
3. Group the attributes and calculate the utility function using $U = \sum [w_i f(S_i)]$
4. Translate the logic expressions in CNF into a set of inequalities using unit matrix.
5. Establish the objective function to be maximize using the Utility Units (u)
6. Use the LP algorithm in the framework to optimise the objective function
7. Convert the evaluated value after the optimisation process to percentage aka $\{u\}$
8. Identify all the Risk factors in the problem domain and build statistical model;
9. Calculate the Predictive Percentage Risk (PPR) by adding: $P(T) \geq P(t) + \{u\}$
10. Map the optimal X values from the optimisation process with the attributes.

In this section, is an illustration of the operation of the proposed Prognosis CVD framework: In this example the Prognosis framework is used to determine the predictive percentage risk (PPR) of CVD disease for the same participants used in Chapter 5. The first part of the framework is to model the CVD disease using class diagram and express the measurable risk factors or attributes in CMAUT formalization. The second step is to calculate the utility unit for each attribute in the expressions using the utility function formulae. Since the organs in the CVD are complementary the total utility unit is the arithmetical sum of the individual utility unit, as shown in the objective function expression (7.0). The objective function is optimised subject to the unit matrix Table 7.1 below, where 1 represents the attribute value measured and 0 indicates the measured attribute but not included in the inequality matrix.

$$Z = \sum_i^n (UH \ XH + UPH \ XPH + UL \ XL + UM \ XM + UT \ XT) \quad (7.0)$$

Table 7.1: Shows attributes values for organs

Utility Attrib	XR	XB	XPH	XH	XM	XT
HB	1	0	0	0	0	0
BMI	0	1	0	0	0	0
BPL	0	0	1	0	0	0
BPH	0	0	1	0	0	0
HDL	0	0	0	1	0	0
MAP	0	0	0	0	1	0
TC	0	0	0	0	0	1

In the CMAUT prognosis framework, the weight allocated to each attribute is assessed using the binary logistic regression in SPSS. The measureable attribute values were entered into the SPSS to determine the beta coefficient values that are used for the computation of the initial clinical absolute risk $\{u\}$. The Location parameter (μ) was calculated by inputting the measureable and non-measurable attribute values into SPSS to determine the beta coefficient of each attribute. Then for the computation of the predictive time factor $P(t)$, a statistical structured regression equation $y = a + \beta_1 X_1 + \dots + \beta_n X_n$ was created using the binary logistic regression technique in SPSS. The results of the clinical absolute risk $\{u\}$ and the factor $P(t)$ were incorporated into the CMAUT algorithm and programmed in MATLAB to compute the PPR value for participant been hypertensive in 10 years.

As part of the illustration, two sets of simulations were conducted using the two Prognosis CMAUT models 1 and 2. The first part of the simulation is to determine the clinical absolute risk $\{u\}$ and the second part of the simulation is the computation of the predictive time factor $P(t)$ for 10 years. The arithmetical sum of the clinical absolute risk $\{u\}$ and predictive time factor $P(t)$ gives the Predictive Percentage Risk (PPR) value of the participant in 10 years. The simulation is based on the hypertension CVD scenario and the (HSE, 2006) clinical data discussed and analysed in chapter 3.

- Data for CVD for modelling CMAUT Prognosis Framework

The demographic data, CVD clinical data and the methodology used to develop the Prognosis framework are from HSE, (2006) report and Craig et al., (2006a). The demographical data used are each participant's series number, age sex, ethinda (i.e. ethnic origin) and clinical data were HB, HDL, BMI, TC, HDL BPH, BPL, ECG/LVC and CVD, Diabetic and Smoking. These parameters were selected for the development of the CMAUT prognosis framework because the Framingham equation and web CHD calculators that the output of the Prognosis framework will be benchmarked against use the same parameters, see Table 5.2.

To illustrate the operation of the prognosis framework models 1 and 2, a participant from the HSE, (2006) data sheet was used see Appendix 3.3 and 3.4. The demographic and CVD clinical data of the select participants, which are Age, Sex, HB, HDL, BMI, TC, HDL BPH, BPL, ECG/LVC and CVD, Diabetic and Smoking were entered into the prognosis framework.

First the framework computes the initial absolute percentage risk (APR) aka $\{u\}$ of the participant been clinical hypertensive. The second part of the prognosis framework, uses the same participant's data to determine the total predictive percentage risk of the participant in 10 years' time. The framework also indicates the attributes in the combinatorial that has the optimal utility values and maps the output variables to their respective attributes. The same simulation exercise was repeated for both model 1 and 2 using the selected 3654 participants' data for the purpose of comparison. The PPR results of the first 10 participants are shown Tables 7.2A and 7.3A while the computed results of the 30 participants and results of the entire groups are in the Appendix 7.

- Domain scenario used for modelling CMAUT CVD Prognosis Framework

To illustrate the operation of the framework, the CVD scenario discussed in Chapter 3 was used and formulated as follows: the hypertension disease (G_1) “is caused by” high rate of pumping blood by the Heart (H) that “sends” excessive high pressure blood to the Atrial(A) which “send signal to” the Antidiuretic hormone ADH in the Brian (B) to regulate the flow of fluid to the kidneys (K). These organs are complementary because they assist each other in performing their duties. In this example, three participants were selected from HSE, (2006) as shown in Table 7.2.

Simulation: Using the clinical data for the same participants in Table 7.2, the predictive percentage risk (PPR) value of each of the participants in 10 years' time was determined and recorded. The results of the first 30 participants and the entire groups are shown in the Appendix 7.

Table 7.2: The CVD data for the participants used for the illustration and simulations are:

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
42,831,101.00	No	No	35	Men	White	64.50	18.15	135.00	70.50	1.70	92.00	No	4.10	Yes	No	No

7.3 Principle of Modelling CMAUT Prognosis Framework

Diagnosis is the first stage in disease management. Therefore in CIS, Initial clinical absolute Percentage Predictive Risk (APR) is essential because it assists medics in the timely intervention of the cause of the disease (Panagiotakos and Stavrinos, 2006). In this research, initial absolute clinical risk (APR) is defined as the percentage probability of a participant having hypertension in their current state and the clinical data measured as well as recorded. The absolute clinical risk is for diagnosis and it is not based on time.

In prognosis, predictive percentage risk (PPR) is the measure of the likelihood of a patient developing a disease over a specific time period based on the measurable and non-measurable risk factors (Panagiotakos and Stavrinos, 2006). The probability of an individual developing CVD event within a specified time frame depends on risk predictors. The risk factors are Age, Sex, OmpulvalHB, BMI, OmsysvalBPH, BPL, Hdlval1HDL, OmmapvalMAP, Diabete2Diabetic, CholvalTotalCholestrol, Smoking YN, Cvddefined as ExistingCVDYN and EcgbECGYN.

The first subsystem of the CMAUT CVD Prognosis Framework: is the use of the optimisation algorithm in the framework to determine the Initial percentage risk of a participant been hypertensive. It is also used to identify the attribute or set of attributes in the combinatorial organs, which maximises the optimal valuation percentage risk. The Initial clinical absolute percentage risk value (APR) is denoted as (u). Again, the algorithm flags the integer values of the attribute variables listed in the results, which are shown as minimum value 0 or the maximum standard value required for medical decision making.

The second subsystem of the CMAUT CVD Prognosis Framework is used to calculate the 10 years percentage predictive risk values. This was adapted from the Weibull distribution and Framingham equations explained in Anderson et al., (1990) and Anderson, (1991). In these works, it was established that it is feasible to calculate the probability of an event occurring at a specific time when given the location parameter (μ) and dispersion parameter (σ) values. According to Anderson et al. (1991), the probability of an event occurring with respect to time (t) is notated as (u) and it is known as the predictive time factor, which is given as:

$$u = [(\log(t) - \mu)/\sigma] \quad (7.1)$$

In equation (7.1), t is the number of years, that the decision maker wants to predict the CVD event occurring. In this research, the Prognosis Framework is designed to predict the CVD event occurring in 10 years, but in the diagnosis framework the time $t = 0$. Again, in the equation (7.1), μ is known as the location parameter and it is the sum of the products of both measurable and non-measurable CVD risk factors. This is computed by multiplying the risk factors by their corresponding coefficients that were obtained from the statistical binary logistic regression analysis in SPSS. The equation of location parameter μ is shown below:

$$\mu = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6 + \beta_7x_7 + \beta_8x_8 + \beta_9x_9 \\ + \beta_{10}x_{10} + \beta_{11}x_{11} + \beta_{12}x_{12}; + \beta_{13}x_{13}$$

In this equation, β_i are the beta coefficients or the weights obtained from the binary logistic regression analysis and the x_i are the values for the CVD risk factors. These are discussed in the implementation section 7.4.1 and 7.4.2 below.

Scale (dispersion) parameter (σ)

The scale or dispersion parameter (σ) is the arithmetical sum of the constant value θ_0 and the initial variable θ_1 for CHD from the Table of SBP prediction equation coefficients (Anderson1990). The constant and initial values are from the results of various studies conducted using different CVD prediction models by Anderson et al, (1990) and Anderson et al. (1991). According to Odell et al., (1994), $\log(\sigma) = \theta_0 + \theta_1\mu$, where θ_0 is a constant and θ_1 the initial variable for the CHD. In this expression, the dispersion parameter (σ) uses the location parameter μ , which depends on the risk factors specified in section 7.2.

In this research, the values θ_0 and θ_1 for the equation $\log(\sigma) = \theta_0 + \theta_1\mu$, are taken from the SBP prediction equation coefficients Table in Anderson et al., (1991). The SBP values were taken from the prediction coefficient Table because all the binary logistic regression analysis conducted with SPSS, the DBP were not included in the output variables required for the structural equation. For example see “Variables in the Equation” Table 7.3 below. Secondly, the values for CHD were selected because the output from the Prognosis Framework will be compared to the Web based CHD calculators. Lastly, the CHD prognosis is specific to heart disease whereas the CVD is generic for the entire Circulatory system. From the Table the values chosen were $\theta_0 = 0,9145$ and $\theta_1 = 0,2784$.

The equation $\log(\sigma) = \theta_0 + \theta_1\mu$ can also be written in natural log form as

$$\text{Loge}(\sigma) = \exp(\theta_0 + \theta_1\mu) \quad (7.2)$$

Hence taken the natural log of equation (7.2) and using the values from Anderson, (1991) the equation will be: $\sigma = e^{\theta_0 + \theta_1\mu}$ and $\sigma = e^{0.9145 + 0.2784\mu}$

Furthermore, from equation (7.1), the predictive time factor $u = [(\log(t) - \mu)/\sigma]$ and for 10years it will be $[(\log(10) - \mu)/\sigma]$.

- Computation of Predictive Percentage Risk – (PPR)

From Anderson, (1991), it is subsumed that the predicted probability represented as P at the time T when the CVD event occurs is greater than the initial time t for the given values of location (μ) and scale (σ) parameters. Hence the assumption that

$$P(T > t) = P\{(\log(T) - \mu)/\sigma > u\} \quad (7.3)$$

Considering that from equation (7.1) $u = (\log(T) - \mu)/\sigma$ then

$$P(T) > P(t) = P\{(\log(T) - \mu)/\sigma \geq u\}$$

From the equation (7.3) above the predicted probability P for 10 years will be

$$P(T) \geq P(t) + \{u\} \quad (7.4)$$

In summary, from the expression (7.4), the predictive risk at time $P(T)$ is greater than at the time $P(t)$. Therefore the $P(T)$, the predictive time in 10years is equal to is the sum of the current percentage risk value and the u value in equation (7.1) (Anderson et al (1991).

Therefore the CMAUT prognosis framework was designed using the following concept:

1. determine the initial clinical absolute risk (APR) or (u) using the CMAUT framework;
2. Add the outcome to the calculated predictive time value $P(t)$
3. The PPR or $P(T)$ in 10 years is the sum of the (APR) and $P(t)$ for each participant.

7.4 Implementation of CMAUT CVD Prognosis Framework – Model 1 and 2

From the HSE, (2006) report, the demographic and clinical data of all participants who are over 16 years and provided full CVD data were selected. The Prognosis Model 1 was designed using the 4316 participants who are over 16 years old because they constitute the core sample data for the HSE, (2006) survey report (Craig, et al, 2006a) (Craig, et al, 2008).

The Prognosis Model 2 is from the data set of the 3654 participants who are over 30 years old and supplied all the requisite CVD data. This is because all the Web CHD risk calculators and the Framingham equations examined in this research were designed for users who are over 30 years old. Therefore adults who are over 30 year's old group, which constitute model 2 was selected (Chuang et al., 2007). The two prediction models built in this research are based on the “Variables in the Equation” from the SPSS logistic regression.

7.4.1. Implementation of CMAUT CVD Prognosis Framework – Model 1

The demographic and clinical data of the 4316 participants who are over 16 years old were used for the design and development of the CMAUT Prognosis model 1. First, the values of the measureable attribute of the entire 4316 participants were input into SPSS and the binary logistic regression conducted. The output from the SPSS gave the values of the beta coefficients for each measureable attributes, which are -10.26, (HBI) 0.211, (BMI) 0.077, (BPH) -0.285, (HDL) 0.200, (MAP) 0.335 (TC) 0.0766. These values listed in the “Variables in the Equation” Table 5.3 were used to form the equation and weights needed to calculate the utility unit of each of the measureable attribute. $Y = -10.261 + 0.211xHB + 0.077xBMI + -0.285xBPH + 0.200xHDL + 0.335xMAP + 0.076xTC$. The formula was used to compute the utility unit and the objective function shown in (7.5) was formulated using the results from the utility unit calculation.

$$\sum_i^n ((-10.261) + 6.75X_R + 3.714X_V + (-3.56)X_{PH} + (-10)X_D + (0)X_M + 0.304X_T) \quad (7.5)$$

In the CMAUT prognosis framework, the measureable attribute values are used for the computation of the initial clinical absolute risk (APR) or $\{u\}$. The CVD problem was presented in LP format as an objective function as shown in expression (7.5) and optimised subject to the constraint unit matrix. The algorithm was implemented in MATLAB program and executed as follows.

Implementation Part 1: execution of CMAUT model 1 and results of initial absolute risk $\{u\}$.

During the execution of the MATLAB CVD program, the following conditions are checked:

1. If the participant is male then $HDL1 = 5.3$, else for female $HDL1 = 5.4$.
2. If the participant is diabetic (Yes) then $BPH1 = 130$; $BPL1 = 80$ else if not diabetic $BPH1 = 140$; $BPL1 = 90$.

The aim of the optimisation algorithm in the framework is to find the attribute(s) in the combinatorial organs that has an overall utility unit that maximizes the utility value to be retrieved for primary healthcare investigation. Therefore the LP optimisation algorithm in MATLAB determines the optimal valuation attribute and the maximum value. The solution indicates that the maximum value or the initial clinical absolute risk (u) is 19.729 as shown in Figure 7.3 below.

7.4.2 Determination of CVD PPR using CMAUT Framework Model1

The second part of the simulation extends the first CMAUT program and incorporates the formulae for predicting the CVD risk in 10 years. In this second part, the measurable and non-measurable CVD risk predictors as well as the Weibull predictive time factor were used to calculate $P(t)$ and add it to the initial clinical absolute risk (u). For both diagnosis and prognosis the weight allocated to each attribute were assessed using the logic regression method. For prognosis, the weights are the values of the beta coefficient for each risk predictor that was calculated with the aid of the binary logistic regression in SPSS. The regression equation used is $y = a + \beta_1x_1 + \dots + \beta_nx_n$.

- Part 2 of the CMAUT Prognosis model 1 time based calculation and results

To determine the Location (μ) and Scale (σ) parameters, the statistical modelling technique was used for the computation of the predictive time factor $P(t)$. The statistical modelling concept uses the parametric regression logistic methods for developing the formula. The formula estimates the probability of the disease occurring when the level of risk factors are given. It allows the use of standard accelerated failure time model for the computation of different duration of follow-up that matches the Weibull distribution (Anderson et al., 1990).

Location parameter μ

To determine the Location parameter the values of the measurable and non-measurable attributes of the 4316 participants in group 1 were input into SPSS and the binary logistic regression conducted. The procedure used is shown in Figure 7.1 below and the attributes used are Age, Sex, OmpulvalHB, BMI, OmdiavalBPL, OmsysvalBPH, Hdlval1HDL, OmmapvalMAP, Diabete2Diabetic, CholvalTotalCholestrol, Smoke4SmokerYN, CvddefExistingCVDYN, and EcgbECGYN. The generic equation for the Location parameter is below and the x_i represents each of the thirteen (13) risk factors.

$$\mu = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6 + \beta_7x_7 + \beta_8x_8 + \beta_9x_9 + \beta_{10}x_{10} + \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13};$$

The expression takes in consideration the entire measurable and non-measurable attributes also known in epidemiological prediction model as CVD risk factors. The CVD risk factors were entered into the SPSS software and the output from the SPSS gave the beta coefficient value for all the attributes, except OmdiavalBPL. The beta coefficient values of the risk factors are listed in the “Variables in the Equation” Table 7.3. The beta coefficients were used as the weights in the equation to calculate the Location parameter.

Procedure for determining the binary logistic Regression in SPSS is as follows:

- Step 1: Analyse → Regression → Binary Logistic
- Step 2: Select bp1 as Dependent variable and move it to the text box
- Step 3: Select: Age, Sex, OmpulvalHB, BMI, OmdiavalBPL, OmsysvalBPH, Hdlval1HDL, OmmapvalMAP, Diabete2Diabetic, CholvalTotalCholestrol, Smoke4SmokerYN, CvddefExistingCVDYN, EcgbECGYN and move them to Covariate text box on the right
- Step 4: Click on Options,
- Step 5: Select classification plot, Hosmer and Lemeshow Test goodness fit, correlation estimate, CI of 95%
- Step 6: Click on OK

Figure 7.1: Procedure for determining Logistic Regression in SPSS

From the statistical regression analysis conducted for model 1, the coefficient values in the Table 7.3 were identified from using the HSE, (2006) clinical data. Therefore the expression below has twelve (12) attributes instead of the thirteen (13) without the OmdiavalBPL.

$$\mu = \beta_0 + \beta_1age + \beta_2sex + \beta_3hb + \beta_4bmi + \beta_5hbp + \beta_6hdl + \beta_7map + \beta_8dia + \beta_9tc + \beta_{10}smk + \beta_{11}CVD + \beta_{12}ecg;$$

Table 7.3 Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	Age	.014	.007	4.476	1	.034	1.014
	Sex(1)	-.141	.201	.492	1	.483	.868
	OmpulvalHB	.300	.299	1.008	1	.315	1.350
	BMI	.108	.022	24.878	1	.000	1.115
	OmsysvalBPH	-.429	.448	.918	1	.338	.651
	Hdlval1HDL	.064	.267	.058	1	.810	1.066
	OmmapvalMAP	.490	.447	1.198	1	.274	1.632
	Diabete2Diabetic	-1.644	.376	19.074	1	.000	.193
	CholvalTotalCholestrol	.043	.088	.241	1	.623	1.044
	Smoke4SmokerYN	-.215	.243	.782	1	.377	.806
	CvddefExistingCVDYN	10.096	1.018	98.413	1	.000	24250.272
	EcgbECGYN	-1.059	.199	28.270	1	.000	.347
	Constant	-18.266	1.505	147.256	1	.000	.000

a Variable(s) entered on step 1: Age, Sex, OmpulvalHB, BMI, OmsysvalBPH, Hdlval1HDL, OmmapvalMAP, Diabete2Diabetic, CholvalTotalCholestrol, Smoke4SmokerYN, CvddefExistingCVDYN, EcgbECGYN.

From Table 7.3, the constant β_0 value is -18.266 and all the β_i values are the beta coefficients for each of the CVD risk factors from the binary logistic regression analysis. Below is the final expression when all the β_i values obtained from the SPSS analysis are put into the Location parameter μ formula for CMAUT model 1:

$$\mu = -18.400 + 0.014age + 0.142sex + 0.300hb + 0.108bmi - 0.430hbp + 0.063hdl + 0.491map + -1.643dia + 0.043tc + -0.215smk + 10.093CVD + -1.058ecg.$$

For the computation of Location parameter μ , the female is denoted as 1, while male is 0. When the participant is diabetic then the value is 1 else 0, and if the participant is smoking then the value is 1 else 0. Again, when the participant had CVD or ECG before then the value is 1 else 0. These risk factor values are put into the Location parameter (μ) formula above to calculate the predictive percentage risk (PPR) as shown in Figure 7.13 below:

$$\begin{aligned}\mu &= -18.400 + 0.014(60) + 0.142(1) + 0.300(34) + 0.108(13.20) - 0.430(122.50) \\ &\quad + 0.063(1.80) + 0.491(100) + -1.643(0) + 0.043(5.02) + -0.215(1) \\ &\quad + 10.093(0) + -1.058(1).\end{aligned}$$

$$\mu = -18.400$$

$$\sigma = e\theta_0 + \theta_1\mu \quad \text{hence } \sigma = e^{0.9145 + 0.2784\mu}$$

$$\sigma = e\theta_0 + \theta_1\mu = 0.0149$$

$$U = \{(\log(T) - \mu)/\sigma\} = \{(\log(10) - \mu)/\sigma\}$$

$$P(T) \geq \text{abs}\{P(t) + \{u\}\}$$

Figure 7.3: Computation of PPR for CMAUT Prognosis model 1

The Figure 7.3 is the code, which depicts how the SPSS results and the predictive time based calculation, which is then incorporated into MATLAB program in Figure 7.4. The output screen for one of the participant from the HSE, (2006) is shown in Figure 7.5 below. The screen depicts the measured attribute values of the participant which was converted into utility unit using the formula 5.1 and the standard parameters in Table 5.2. The optimisation algorithm is used to find the attribute(s) in the combinatorial organs that has an overall utility unit that maximizes the utility value to be retrieved for primary healthcare investigation.

- Determination of CVD PPR value using the Model1 CMAUT framework

The optimisation algorithm was written in MATLAB and used to simulate the CVD data of each participant. The MATLAB program in Figure 7.4 uses each participant's data and computes the 10 years predictive percentage risk (PPR). The algorithm in the MATLAB code shows the arithmetical sum of the $\{\text{abs } P(t)\}$ and $\{\text{abs } (u)\}$, which is the sum of the absolute percentage risk (u) and the time based predictive percentage risk ($P(t)$). The first and second parts of the CMAUT Prognosis model 1, were implemented in MATLAB. The algorithm and formulae used are coded in MATLAB are as follows:

```

const = -18.266;
n = length(fage);
mu_val = zeros(1,n);
u_val = zeros(1,n);
pr_val = zeros(1,n);
for i = 1:n
%   tc_div_hdl = ftc(i)/fhdl(i);

    mu_val(i) = const + 0.014*fage(i) + (-0.141*fsex(i)) + 0.300*fhb(i)+
        0.108*fbmi(i) + (-0.429*fbph(i)) + 0.064*fhdl(i) +
        0.490*fmap(i) + (-1.644*fdia(i)) + 0.043*ftc(i) +
        (-0.215*fsmk(i)) + (-1.059*fecg(i)) + 10.096*fcvd(i);
    log_sig = 0.9145 - 0.2784 * mu_val(i); % theta one - theta two * mu
    sig_val = exp(log_sig);
    u_val(i) = (log(10) - mu_val(i))/sig_val;
    if i == 18
        gg = dpts(i);
    end
    pr_val(i) = dpts(i)+abs(u_val(i));
end
pr_val = pr_val';
figure(1)
plot(pr_val);

```

Figure 7.4: MATLAB code for CMAUT Prognosis model 1

In part 2 of simulation process, the optimisation algorithm in the CMAUT framework uses the entire CVD predictors and the 10 years risk predictive formulae. When the CMAUT time based framework in MATLAB is executed, the results appear in the GUI in Figure 7.5 below. The GUI shows the initial Absolute Percentage Risk (APR) and the Predictive Percentage Risk (PPR) of the named participant in 10 years' time. The output also flags the attributes in the combinatory, which are needed for the management of the participant's CVD disease.

In the GUI, the results after optimisation and the optimal integer values are $X_1 = 50$, $X_2 = 26$, $X_3 = 1$, $X_4 = 1$, $X_5 = 100$ and $X_6 = 6$. As shown in the GUI solution Figure 7.5, the percentage risk is 19.729 % and the optimal values of the attributes are $X_1 = 50$, $X_2 = 26$, and $X_5 = 100$, which are the requisite data for further investigation. When the output variables are mapped to the input attributes the results are X_1 (HB), X_2 (BMI), X_5 (MAP) and PPR value is the arithmetical sum of the $\{abs P(t)\}$ and $\{abs(u)\}$, which is 20.0112 %.

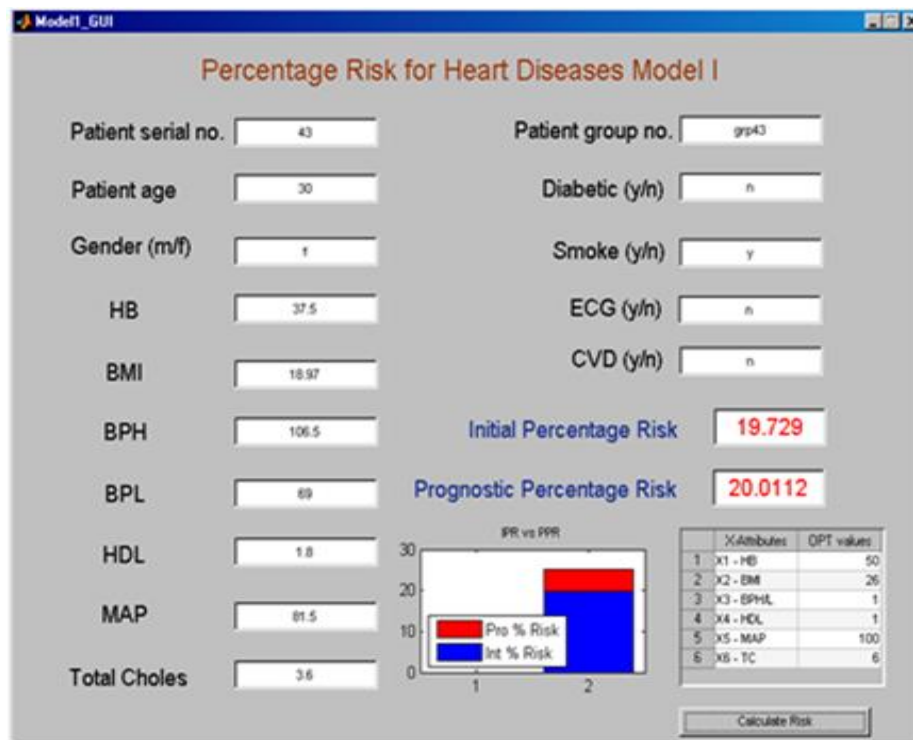


Figure 7.5: Output screen of CIS_CMAUT framework with predictive time for simulation 2

- Simulation Results, Tables and Figures for CMAUT Prognosis framework model1

The Tables and Figures in this section are the results of entering the demographic and clinical data of each of the selected 3654 participants into the CVD CMAUT Prognosis framework model 1. The PPR results from the CVD Prognosis framework model 1 are compared with the risk results from the Web based CVD calculators and Framingham equations in chapter 9.

Table 7.3A is the raw data of the first 10 participants from the list of 3654 participants. Table 7.3B at the end of the Thesis contains the results of the first 30 participants and the group results are in the Appendix Table 7.3C in electronic format. Table 7.4A contains the results of the computed 10 years PPR values and the attribute variable values of each of the first 10 participants from the Model I 3645 data sets. Table 7.4B, at the end of the Thesis is the results of the 30 participants and the group is in Appendix Table 7.4C in electronic format.

The Table 7.5A shows the results of the computation of TPR, FPR, LRP and LRN for the CMAUT CVD Prognosis framework model 1. The 10 years PPR values are for the first 10 participants of the 3645 data set. The Table 7.5B, at the end of the Thesis contains the first 30 participants' results and the group is in the Appendix Table 7.5C in electronic format.

MATLAB Model I for 3645 participants in the category of over 16 years old.

Table 7.3A: the raw data of the first 10 participants used in chapter 7:

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.0	13.20	122.50	88.50	1.80	100.0	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.0	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
71,831,101.00	No	No	66	Women	White	89.0	14.32	159.00	70.00	1.90	99.50	No	6.90	No	No	Yes
34,031,101.00	No	No	84	Women	White	48.5	16.17	112.00	63.50	2.20	80.00	No	5.00	Yes	Yes	Yes
72,604,102.00	No	No	59	Women	White	36.5	16.19	109.50	73.00	2.00	85.00	No	6.00	No	No	No
13,008,101.00	Yes	Yes	50	Women	White	43.0	16.65	117.00	74.00	1.70	88.50	No	6.00	Yes	Yes	Yes
39,139,101.00	No	No	34	Women	White	48.0	16.81	102.00	54.00	1.80	70.00	No	6.50	Yes	No	No
47,856,102.00	No	No	51	Women	White	44.0	16.85	100.50	56.50	1.90	71.00	No	5.10	No	No	No
37,710,101.00	No	No	61	Women	White	43.0	17.43	120.00	77.00	1.20	91.50	No	5.50	No	No	No

Table 7.4A Predicative Percentage Risks for 10 years and attribute variable values for the first 30 participants (from Model I 3645 data sets)

Pserial no.	Grp	Bp1	Age	Sex	X1	X2	X3	X4	X5	X6	%PR
13,956,102.00	No	No	60	Women	50.0	25.5	0.02	0.00	50.00	5.22	14.9
63,535,102.00	Yes	Yes	30	Women	45.8	25.9	4.15	0.00	100.19	4.28	16.3
71,831,101.00	No	No	66	Women	0.0	25.5	140.00	0.00	100.00	0.00	17.1
34,031,101.00	No	No	84	Women	50.0	25.5	0.00	0.00	100.00	5.30	19.0
72,604,102.00	No	No	59	Women	50.0	25.5	0.00	0.00	100.00	0.00	19.2
13,008,101.00	Yes	Yes	50	Women	50.0	25.5	0.00	0.00	100.00	0.00	17.3
39,139,101.00	No	No	34	Women	50.0	25.5	0.00	0.00	100.00	0.00	21.7
47,856,102.00	No	No	51	Women	50.0	25.5	0.01	0.00	100.00	5.23	22.2
37,710,101.00	No	No	61	Women	50.0	25.5	0.00	0.00	100.00	0.00	15.6
54,256,101.00	No	No	31	Women	50.0	25.5	0.00	0.00	100.00	5.30	14.4

Table 7.5A Calculation of TPR, FPR, LRP, and LRN, for the MATLAB Model I PPR for 10 years for the first 30 participants (from Model I 3645 data sets)

Pserial no.	Grp	Bp1	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRN
13,956,102.00	No	No	60	Women	15.6	0	1	1	1.000	3278.7	0
63,535,102.00	Yes	Yes	30	Women	16.3	0	1	1	0.999	1639.3	0
71,831,101.00	No	No	66	Women	17.7	0	1	1	0.999	1092.9	0
34,031,101.00	No	No	84	Women	19.0	0	1	1	0.999	819.7	0
72,604,102.00	No	No	59	Women	19.8	0	1	1	0.998	655.7	0
13,008,101.00	Yes	Yes	50	Women	21.0	0	1	1	0.998	546.4	0
39,139,101.00	No	No	34	Women	19.4	1	0	0.997	0.998	545.0	0.0027
47,856,102.00	No	No	51	Women	17.2	1	0	0.995	0.998	543.5	0.0055
37,710,101.00	No	No	61	Women	18.3	0	1	0.995	0.998	465.8	0.0055
54,256,101.00	No	No	31	Women	22.2	0	1	0.995	0.998	407.4	0.0055

Figure 7.3 shows the graph of the computation of the 10 years PPR value of each of the 3645 participants against their individual Participant Identification Number (PIND). This data in Table 7.4C were used to plot the graph is shown in Figure 7.3.

Figure 7.4 shows the prediction accuracy graph, which is the computation of True Positive Rate against the False Positive Rate of each of the 3645 participants. This data used for plotting the Area under the Curve graph in Figure 7.4 is from Table 7.5C. The AUC for model 1 is calculated, by using the summation of all the PPR data points applying the trapezoidal method. The AUC area is obtained by subtracting the sum of all the PPR data points from the sum of all the diagonal reference data points.

The Figure 7.5 shows the graph of the discriminatory ability of the CMAUT model 1. This was constructed by first calculating the sensitivity and selectivity of each of the 3645 participants and recorded in Table 7.5C for the results of the calculation of TPR and FPR for model 1. The graph Figure 7.5 of the sensitivity and selectivity are plotted against the recommended criterion, using the NICE, (2006) criterion of 20 %. The interception and the degree of accuracy are discussed in chapter 9.

Likelihood ratio is used to determine the performance accuracy of the model in Table 7.5C. The procedure used to calculate the value of the Likelihood ratio $LR+ = (TPR/1-TNR)$ and the Likelihood ratio $LR- = (1-TPR/TNR)$ for all the 3645 participants. The Table 7.5C shows the results of the calculation of the $LR+$ and the $LR-$ for model 1. Figure 7.6 shows the graphic of the performance accuracy of the PPR of each participant's value from the Prognosis framework model 1. The graph in Figure 7.6 was constructed by plotting all the positive and negative Likelihood ratio values on the Y-axis and PIND of each participant on the X-axis and discussed in chapter 9.

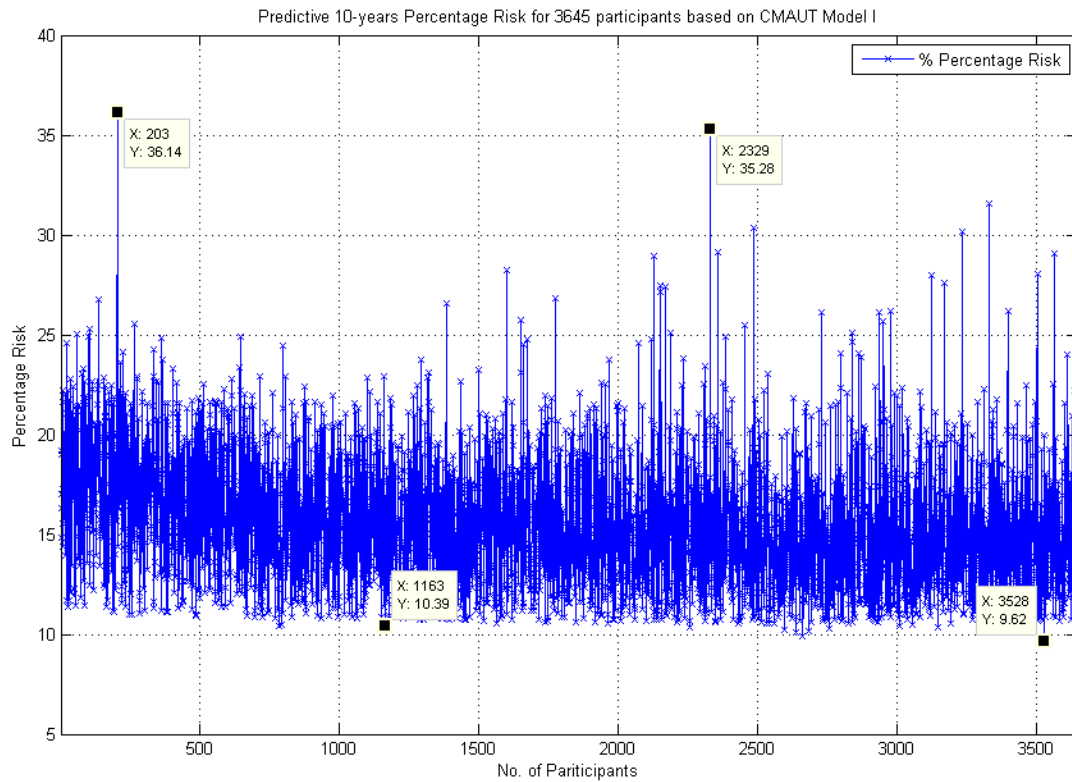


Figure 7.3: Predictive 10-years percentage risk for 3645 participants based on CMAUT Model – I

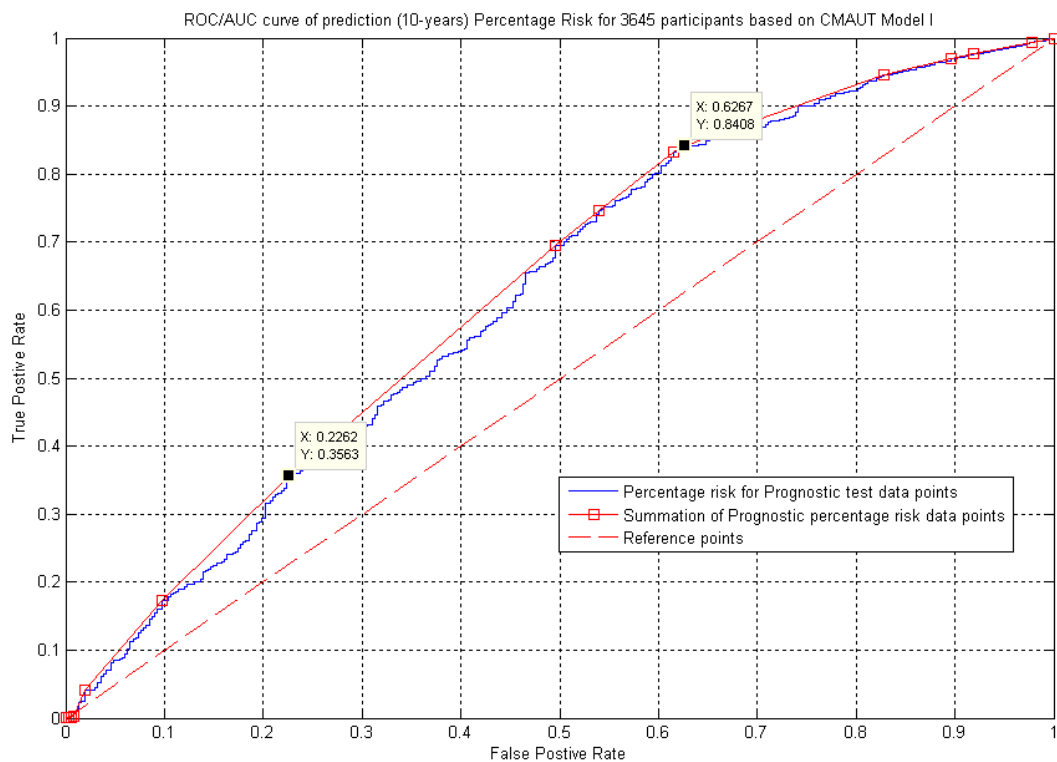


Figure 7.4: ROC/AUC curve of CMAUT Prognosis Risk Model – I

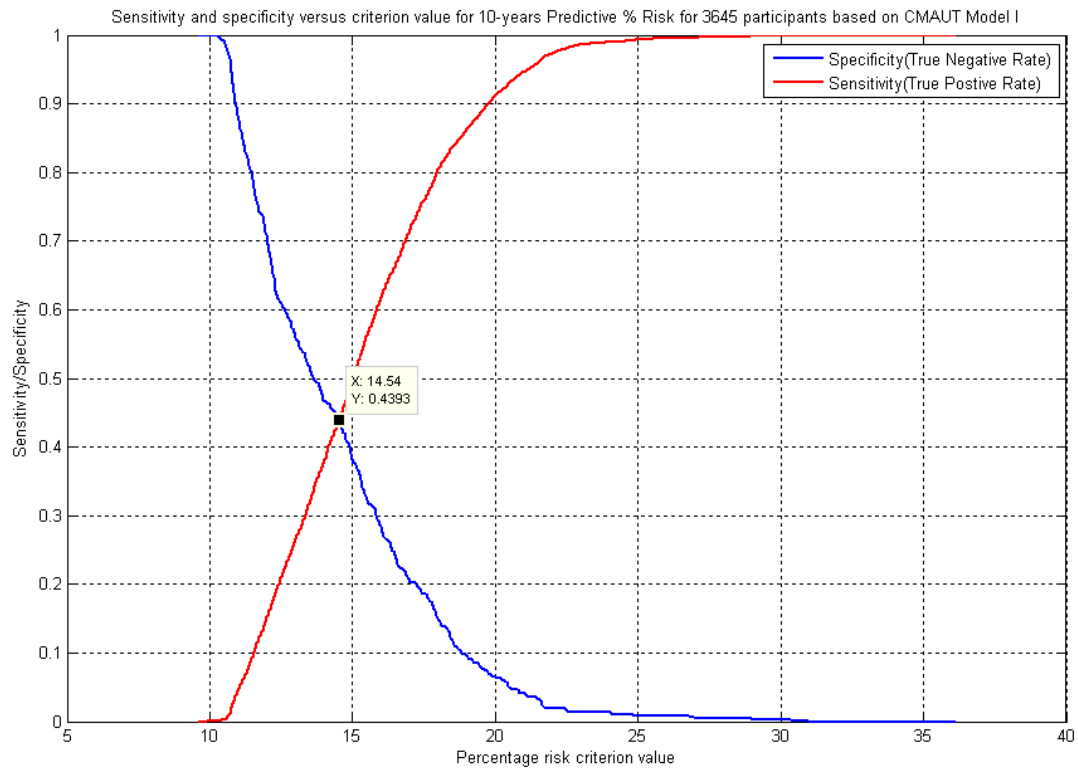


Figure 7.5: Sensitivity and specificity curve of CMAUT Prognosis Risk Model – I

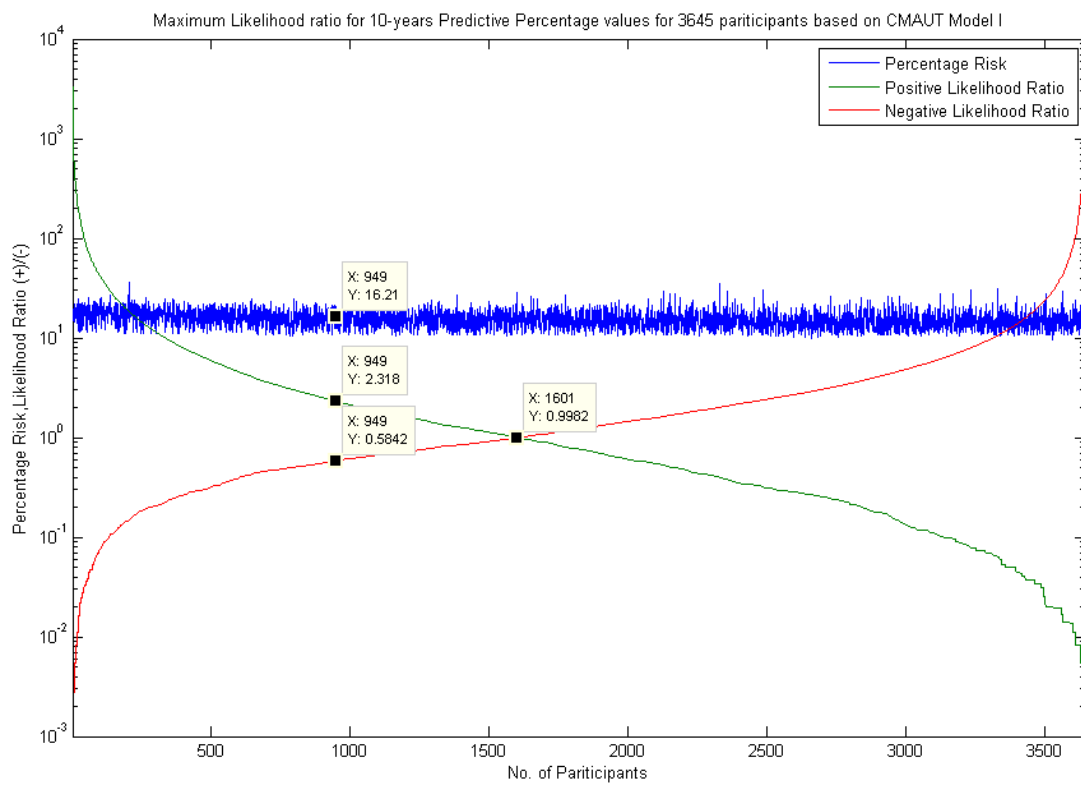


Figure 7.6: Likelihood ratio curve of CMAUT Prognosis Risk Model – I

7.4.3 Implementation of CMAUT CVD Prognosis Framework – Model 2

The demographic and clinical data of the 3645 participants who are over 30 years old were used for the design and development of the CMAUT prognosis model 2. First, the values of the measureable attribute of the entire 3645 participants were input into SPSS and the binary logistic regression conducted. The output from the SPSS gave the beta coefficient values for each measureable attribute, which are OmpulvalHB = 0.174, BMI = 0.076, OmsysvalBPH = -0.229, Hdlval1HDL = 0.216, OmmapvalMAP = 0.279, CholvalTotalCholestrol = 0.058 and the constant value = -10.076. The values listed in the “Variables in the Equation” Table 5.5 were used to form the equation below and the same values used as weights to calculate the utility unit of each of the measureable attribute.

$$y = -10.076 + 0.174x_1 + 0.076x_2 + -0.229x_3 + 0.216x_4 + 0.279x_5 + 0.058x_6.$$

The formula used to calculate the utility unit is the expression (5.1) and the objective function formed from the results is in expression (7.4) below.

$$Z = \sum_i^n ((-10.076) + 5.568X_R + 3.665X_V + (-2.86)X_{PH} + (-10.8)X_D + (0)X_M + (-0.232)X_T) \quad (7.4)$$

In CMAUT prognosis framework the measureable attribute values are used for the computation of the initial clinical absolute risk (APR) or $\{u\}$. The CVD problem was presented in LP format as an objective function and optimised subject to the constraint unit matrix. The algorithm was implemented in MATLAB program.

- Implementation part 1:- execution and results of CMAUT model 2 for APR $\{u\}$

During the execution of the MATLAB program, it checks the following conditions that:

1. If the participant is male then HDL1 = 5.3 else for female HDL1 = 5.4.
2. If the participant is diabetic (Yes) then BPH1 = 130; BPL1 = 80 else if not diabetic BPH1 = 140; BPL1 = 90.

The LP optimisation algorithm written in MATLAB is used to determine the optimal maximum valuation attribute value. The solution indicates that the maximum value, which is the initial clinical absolute risk (u) shown in Figure 7.8 as 18.007%.

7.4.4 Determination of CVD PPR using CMAUT framework Model 2

The second part of the simulation extends the CMAUT program from the first part and incorporates the formulae for predicting the risk in 10 years. In this second part, the measurable and non-measurable CVD risk predictors as well as the Weibull predictive time factor $P(t)$ are calculated and added to the initial clinical absolute risk (u).

For prognosis, the weights are the values of the beta coefficient for each risk predictor as calculated using the binary logistic regression in SPSS. The regression equation used is: $y = a + \beta_1x_1 + \dots + \beta_nx_n$:

- Part 2 of the CMAUT Prognosis model 2 time based calculation and results

To determine the Location (μ) and Scale (σ) parameters required for the computation of the predictive time factor the statistical modelling technique was used. The statistical modelling concept uses parametric regression, which is logistic method for developing the formula. The formula estimates the probability of the disease occurring when the values of risk factors are given. It allows the use of standard accelerated failure time model for computation of different duration of follow-up that matches the Weibull distribution (Anderson et al., 1990).

Location parameter μ

The Location parameter was determined by inputting the measureable and non-measurable attribute values of the entire 3645 participants into SPSS software and the binary logistic regression conducted. The procedure used is shown in Figure 7.7 below and the attributes used are Age, Sex, OmpulvalHB, BMI, OmdiavalBPL, OmsysvalBPH, Hdlval1HDL, OmmapvalMAP, Diabete2Diabetic, CholvalTotalCholestrol, Smoke4SmokerYN, CvddefExistingCVDYN and EcgbECGYN. The equation for Location parameter is written below and in it the x_i represents each of the thirteen (13) risk factors.

$$\mu = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6 + \beta_7x_7 + \beta_8x_8 + \beta_9x_9 \\ + \beta_{10}x_{10} + \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13};$$

The Location parameter formula uses the entire measurable and non-measurable attributes the CHD risk factors used in epidemiological prediction model were input into the SPSS software. The output from the SPSS gave the values of the beta coefficients for all the attributes, except OmdiavalBPL. The values listed in the “Variables in the Equation” Table 7.13 were used to form the equation below and also used as the weights to calculate the Location parameter of all the CVD risk factors.

Procedure for determining logistic Regression in SPSS

Step 1: Analyse → Regression → Binary Logistic

Step 2: Select bp1 as Dependent variable and move it to the text box

Step 3: Select: Age, Sex, OmpulvalHB, BMI, OmdiavalBPL, OmsysvalBPH, Hdlval1HDL, OmmapvalMAP, Diabete2Diabetic, CholvalTotalCholestrol, Smoke4SmokerYN, CvddefExistingCVDYN, EcgbECGYN and move them to Covariate text box on the right

Step 4: Click on Options,

Step 5: Select classification plot, Hosmer and Lemeshow Test goodness fit, correlation estimate, CI of 95%

Step 6: Click on OK

Figure 7.7: Procedure for logistic Regression in SPSS

From the statistical regression analysis conducted the beta coefficient values in Table 7.12 were obtained for the HSE, (2006) data for model 2. The expression below has the twelve (12) attributes without the OmdiavalBPL instead of 13 attributes put into the SPSS package.

$$\mu = \beta_0 + \beta_1age + \beta_2sex + \beta_3hb + \beta_4bmi + \beta_5hbp + \beta_6hdl + \beta_7map + \beta_8dia + \beta_9tc + \beta_{10}smk + \beta_{11}CVD + \beta_{12}ecg;$$

Table 7.12 Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	Age	.023	.008	7.619	1	.006	1.023
	Sex(1)	-.111	.212	.276	1	.600	.894
	OmpulvalHB	.241	.317	.578	1	.447	1.272
	BMI	.109	.023	22.052	1	.000	1.115
	OmsysvalBPH	-.352	.475	.548	1	.459	.704
	Hdlval1HDL	.116	.278	.173	1	.677	1.122
	OmmapvalMAP	.419	.475	.778	1	.378	1.520
	Diabete2Diabetic	-1.508	.392	14.824	1	.000	.221
	CholvalTotalCholestrol	.060	.094	.412	1	.521	1.062
	Smoke4SmokerYN	-.063	.265	.057	1	.812	.939
	CvddefExistingCVDYN	10.018	1.021	96.242	1	.000	22419.271
	EcgbECGYN	-1.106	.214	26.784	1	.000	.331
	Constant	-19.228	1.605	143.457	1	.000	.000

a Variable(s) entered on step 1: Age, Sex, OmpulvalHB, BMI, OmsysvalBPH, Hdlval1HDL, OmmapvalMAP, Diabete2Diabetic, CholvalTotalCholestrol, Smoke4SmokerYN, CvddefExistingCVDYN, EcgbECGYN.

In Table 7.12, the constant value of β_0 is -18.266 and all the beta coefficient (β_i) values for the various CVD risk factors from the binary logistic regression analysis are shown. Below is the final expression when all the β_i values obtained from the SPSS analysis are put into the Location parameter μ formula for CMAUT model 2:

$$\begin{aligned}\mu = & -19.228 + 0.023age - 0.111sex + 0.241hb + 0.109bmi - 0.352hbp + 0.116hdl \\ & + 0.419map + -1.509dia + 0.060tc + -0.063smk + 10.018CVD \\ & - 1.106ecg\end{aligned}$$

For the computation of Location parameter μ , the female is given 1, while male is 0. When the participant is diabetic the value given is 1 else 0, and when the participant is smoking then give 1 else 0. Give the participant the value 1, if they have had CVD or ECG before else give them 0. These values are put into the formula to calculate Location parameter μ and predictive percentage risk (PPR) as shown in Figure 7.8 below:

$\begin{aligned}\mu = & -19.228 + 0.023(60) - 0.111(1) + 0.241(34) + 0.109(13.20) - 0.352(122.50) \\ & + 0.116(1.80) + 0.419(100) + -1.509(0) + 0.060(5.02) + -0.063(1) \\ & + 10.018(1) - 1.106(0).\end{aligned}$
$\mu = -18.400$
$\sigma = e^{\theta_0 + \theta_1\mu} \text{ hence } \sigma = e^{0.9145 + 0.2784\mu}$
$\sigma = e^{\theta_0 + \theta_1\mu} = 0.0149$
$U = \{(\log(T) - \mu)/\sigma\} = \{(\log(10) - \mu)/\sigma\}$
$P(T) \geq \text{abs}\{P(t) + \{u\}\}$

Figure 7.8: Computation of Location parameter μ and Predictive Percentage Risk

Figure 7.9 is the MATLAB code, which depicts how the SPSS results and the predictive time based calculations, which were incorporated into the MATLAB program. The output screen for one of the participant from the HSE, (2006) survey is shown in Figure 7.10 below. The output screen shows the measured attribute values of the participant that was converted into utility unit using the formula (5.1) and the standard parameters in Table 5.2.

The optimisation algorithm in the framework is used to determine the attribute(s) in the combinatorial organs that have the overall utility unit that maximises the utility function to be retrieved for CVD investigation.

- Part 2 results – Using the CMAUT time based framework for model 2

The second part of the algorithm for the computation of PPR was written in MATLAB and simulated using the clinical data for each participant. The MATLAB algorithm takes each of the participant's data and computes the 10 years predictive percentage risk PPR value of the participant. The formulae in MATLAB are as follows:

```
const = -19.228;
n = length(fage);
mu_val = zeros(1,n);
u_val = zeros(1,n);
pr_val = zeros(1,n);
for i = 1:n
%   tc_div_hdl = ftc(i)/fhdl(i);
mu_val(i) = const + 0.023*fage(i) + (-0.111*fsex(i)) + 0.241*fhb(i)+...
0.109*fbmi(i) + (-0.352*fbph(i)) + 0.116*fhdl(i) +...
0.419*fmap(i) + (-1.508*fdia(i)) + 0.060*ftc(i) + ...
(-0.063*fsmk(i)) + (-1.106*fecg(i)) + 10.018*fcvd(i);
log_sig = 0.9145 - 0.2784 * mu_val(i); % theta one - theta two * mu
sig_val = exp(log_sig);
u_val(i) = (log(10) - mu_val(i))/sig_val;
pr_val(i) = dpts(i)+abs(u_val(i));
end
pr_val = pr_val';
figure(3)
plot(pr_val);
```

Figure 7.9: MATLAB program for CMAUT Prognosis model 2;

- Simulation results – Using the CMAUT time based framework –model 2

In part 2 of the simulation process, the optimisation algorithm uses all the CVD risk predictors and the 10 years risk predictive $P(T)$ formulae. When the CMAUT time based Prognosis framework is executed the results appear in the GUI in Figure 7.10 below. It comprises the initial clinical percentage risk, just as in simulation part 1. The output also flags the attributes in the combinatority that are needed for the management of the participant's CVD disease.

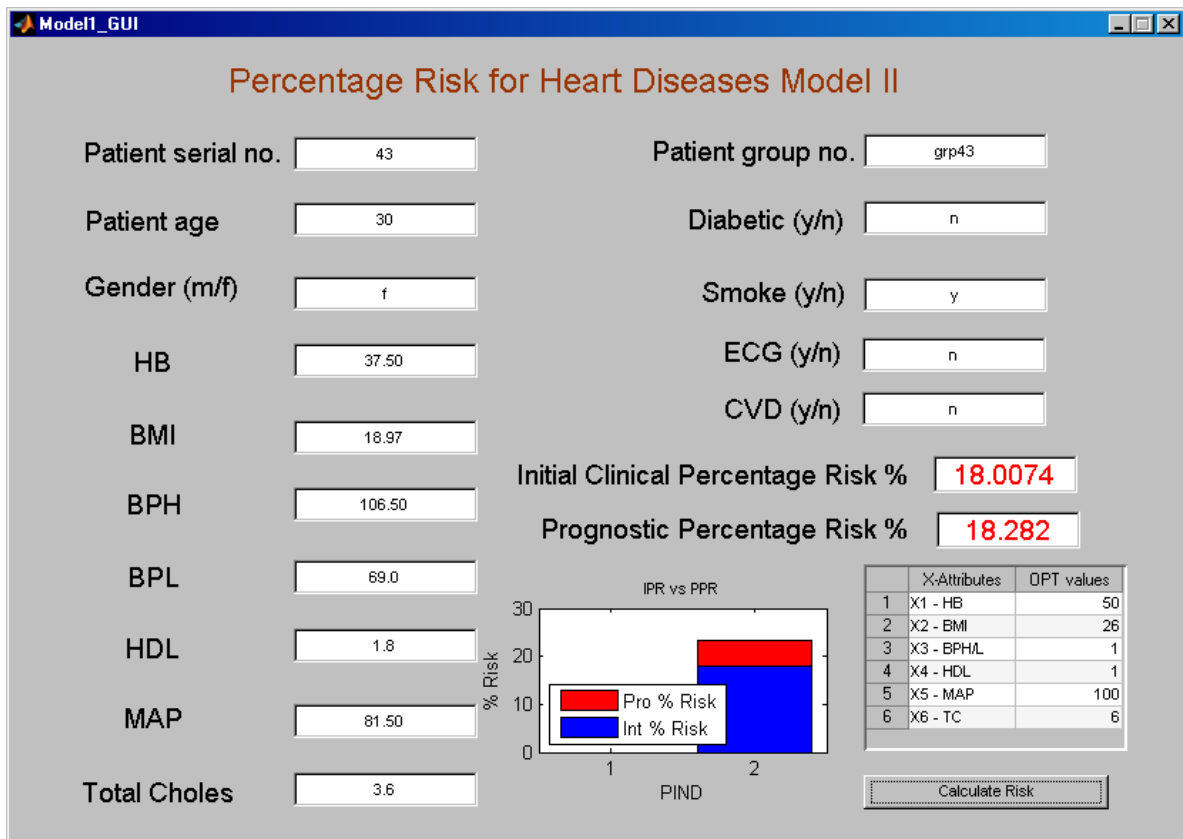


Figure 7.10: Output screen of CIS_CMAUT framework with predictive time for simulation 2.

The results after optimisation are the optimal integer values shown in Figure 7.10 as follows; $X_1 = 50$, $X_2 = 26$, $X_3 = 1$, $X_4 = 1$, $X_5 = 100$ and $X_6 = 6$. The computed PPR value is 18.282 % while the attribute values X_1 (HB) = 50, X_2 (BMI) = 26, and X_5 (MAP) = 100 needs further investigation. From the results in the GUI Figure 7.10, the 10 years $P(T)$ of the participant, which is the arithmetical sum of the $\{abs(u) \text{ and } abs P(t)\}$ is $(18.007 + x) = 18.282\%$. This result denotes that the participant needs attention because the value is close to the 20% recommended by NICE, (2006).

The values in Figure 7.10 are comparable to the values specified by NICE, (2006) as $HB_1 = 50.0$; $BMI_1 = 25.5$; $BPH_1 = 140.0$; $BPL_1 = 90.0$; $HDL_1 = 1.20$; $MAP_1 = 100$; $TC_1 = 5.0$ as used in the CMAUT optimisation algorithm.

- Simulation Results, Tables and Figures for CMAUT Prognosis framework model2:

The Tables and Figures in this section are the results of entering the demographic and clinical data of each of the selected 3654 participants into the CVD CMAUT Prognosis framework model 2. The results from CVD Prognosis framework model 2 were compared with the results from the Web based CVD calculators and the Framingham equations in Chapter 9. The raw data of the first 10 participants from the list of 3654 participants is in Table 7.3A while Table 7.3B contains the data of the first 30 participants at the end of this Thesis and Table 7.3C has the entire group data in electronic format.

Table 7.13A contains the results of the calculated 10 years PPR values and the comparative attribute variable values of each of the first 10 participants from the Model I 3645 data sets. Similarly, Table 7.13B, at the end of the Thesis contains the results of first 30 participants and the results of entire group are in the Appendix Table 7.13C in electronic format.

The Table 7.14A shows the results of the computation of TPR, FPR, LRP and LRN for the CMAUT CVD Prognosis framework model 1 based on the PPR values for the 10 years of the first 10 participants of the 3645 data set. Table 7.14B, at the end of the Thesis contains PPR results of the first 30 participants and the results of the entire group are in the Appendix 7, Table 7.14C in electronic format.

- Table 7.3A contains the raw data of the first 30 participants used in MATLAB Model II for 3645 participants in the category of over 30 years old.

Table 7.13 Predicative Percentage Risks for 10 years and attribute variable values for the first 30 participants Model II

Pserial no.	Grp	Bp1	Age	Sex	X1	X2	X3	X4	X5	X6	%PPR
13,956,102.00	No	No	60	Women	50.0	25.5	0.02	0.00	50.00	5.22	14.79
63,535,102.00	Yes	Yes	30	Women	45.8	25.9	4.15	0.00	100.19	4.28	15.53
71,831,101.00	No	No	66	Women	0.0	25.5	140.00	0.00	100.00	0.00	16.42
34,031,101.00	No	No	84	Women	50.0	25.5	0.00	0.00	100.00	5.30	17.64
72,604,102.00	No	No	59	Women	50.0	25.5	0.00	0.00	100.00	0.00	18.26
13,008,101.00	Yes	Yes	50	Women	50.0	25.5	0.00	0.00	100.00	0.00	16.13
39,139,101.00	No	No	34	Women	50.0	25.5	0.00	0.00	100.00	0.00	20.28
47,856,102.00	No	No	51	Women	50.0	25.5	0.01	0.00	100.00	5.23	20.75
37,710,101.00	No	No	61	Women	50.0	25.5	0.00	0.00	100.00	0.00	15.26
54,256,101.00	No	No	31	Women	50.0	25.5	0.00	0.00	100.00	5.30	14.26

Table 7.14 Calculation of TPR, FPR, LRP, and LRN, for the MATLAB Model II for 10 years for the PPR first 30 participants

Pserial no.	Grp	Bp1	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRM
13,956,102.00	No	No	60	Women	14.79	0	1	1	0.9997	3484.32	0
63,535,102.00	Yes	Yes	30	Women	15.53	0	1	1	0.9994	1745.20	0
71,831,101.00	No	No	66	Women	16.42	0	1	1	0.9991	1162.79	0
34,031,101.00	No	No	84	Women	17.64	0	1	1	0.9989	872.60	0
72,604,102.00	No	No	59	Women	18.26	0	1	1	0.9986	697.84	0
13,008,101.00	Yes	Yes	50	Women	16.13	0	1	1	0.9983	581.73	0
39,139,101.00	No	No	34	Women	20.28	1	0	0.9935	0.9983	577.98	0.0065
47,856,102.00	No	No	51	Women	20.75	1	0	0.9871	0.9983	574.23	0.0129
37,710,101.00	No	No	61	Women	15.26	0	1	0.9871	0.9980	492.07	0.0129

Figure 7.11 shows the graph of the computed 10 years PPR values of each of the 3645 participants against their individual Participant Identification Number (PIND). This data in Table 7.13C was used to plot the graph shown in Figure 7.11. Figure 7.12, is the graph of the computed TPR against the FPR value of each of the 3645 participants. This data used to plot the AUC graph is from Table 7.13C and the AUC area for model 2 was calculated, by the summation of all the PPR data points using the trapezoidal method. The actual area is obtained by subtracting the sum of all the PPR data points from the sum of the of all the diagonal reference data points as discussed in chapter 3.

Figure 7.13 shows the graph of the Prediction Accuracy of CMAUT model 2. This was constructed by plotting the sensitivity and selectivity of each of the 3645 participants from the values in Table 7.13C. The graph of the sensitivity and selectivity against the NICE (2006) recommended 20% criterion and the interception are discussed in chapter 9. Table 7.13C contains the results of the computation of the positive and the negative Likelihood ratio for the entire 3645 participants.

The Figure 7.14 shows the graph of the Prediction Accuracy of the PPR of each participant's value from the Prognosis predictive framework model 2. This is based on the plot all the values of the positive and negative Likelihood ratios on the Y-axis and PIND for each participant on the X-axis and the graph are discussed in chapter 9.

Plots for 3645 data sets for Internet model II

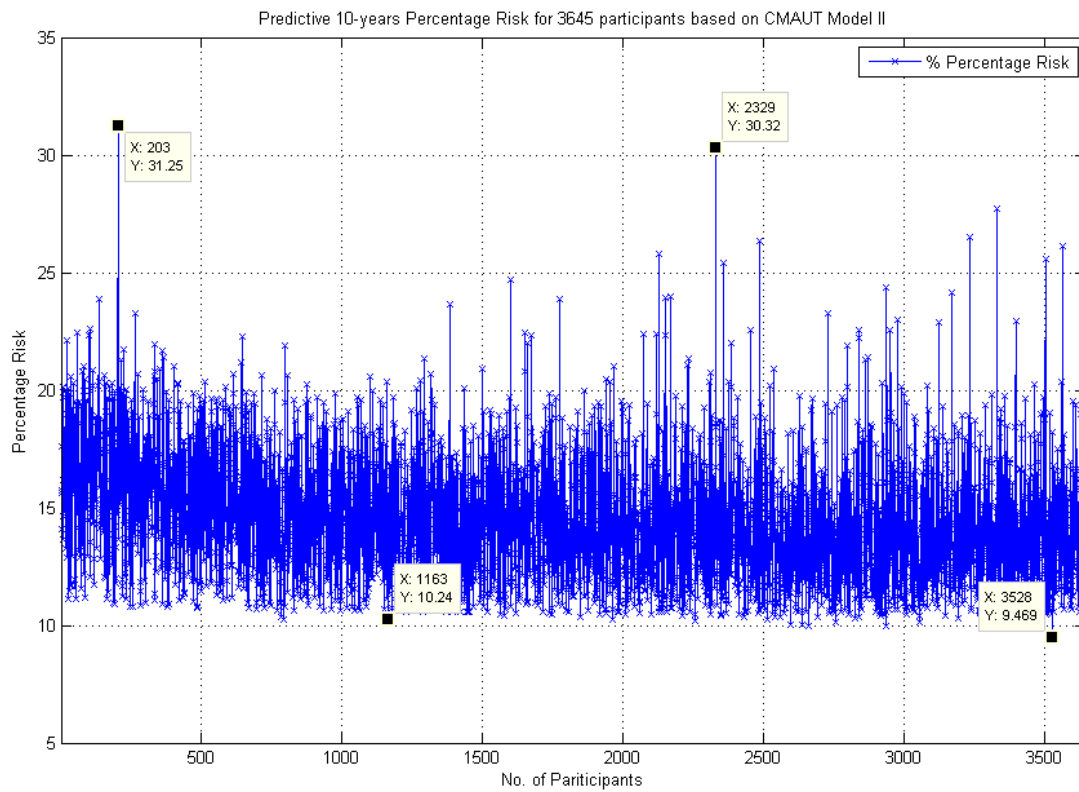


Figure 7.11: Predictive 10-years percentage risk for 3645 participants based on CMAUT Model – II

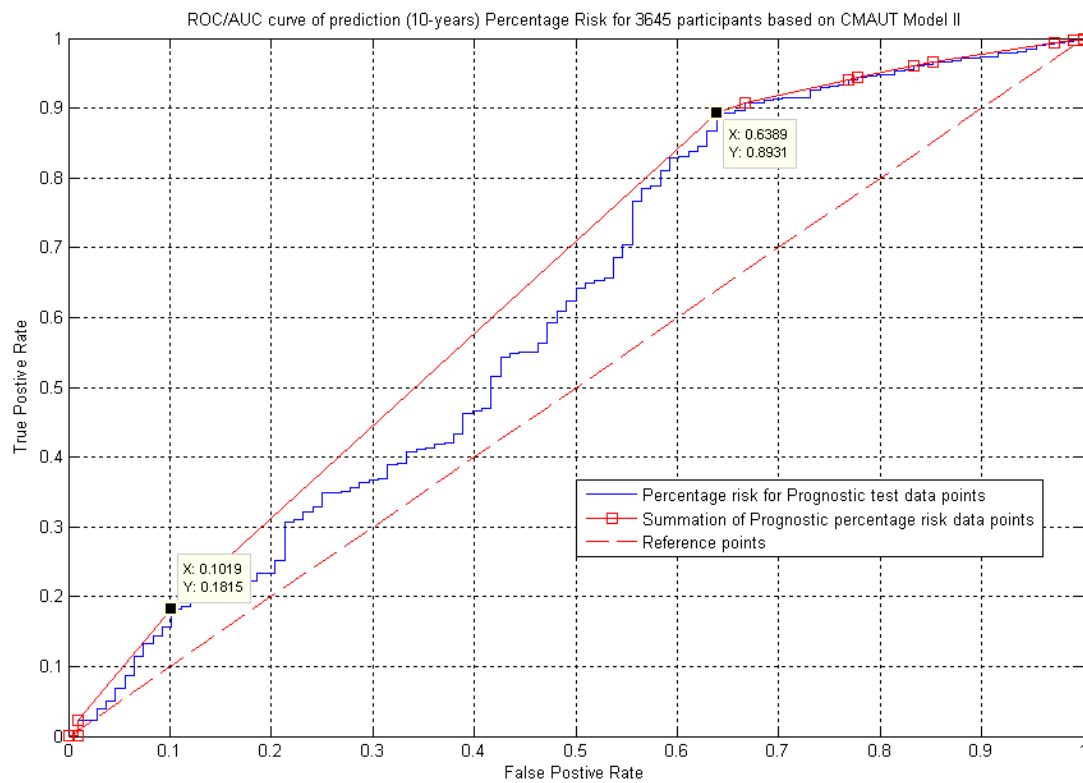


Figure 7.12: ROC/AUC curve for CMAUT Prognosis Risk Model – II

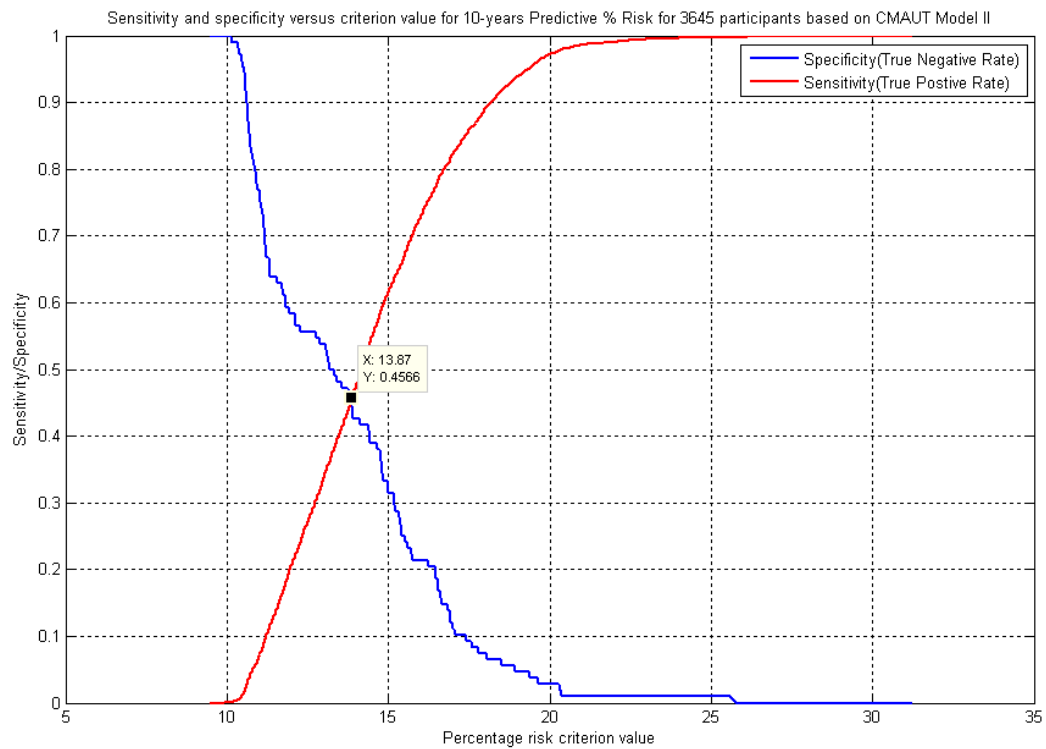


Figure 7.13: Sensitivity and specificity curve of CMAUT Prognosis Risk Model – II

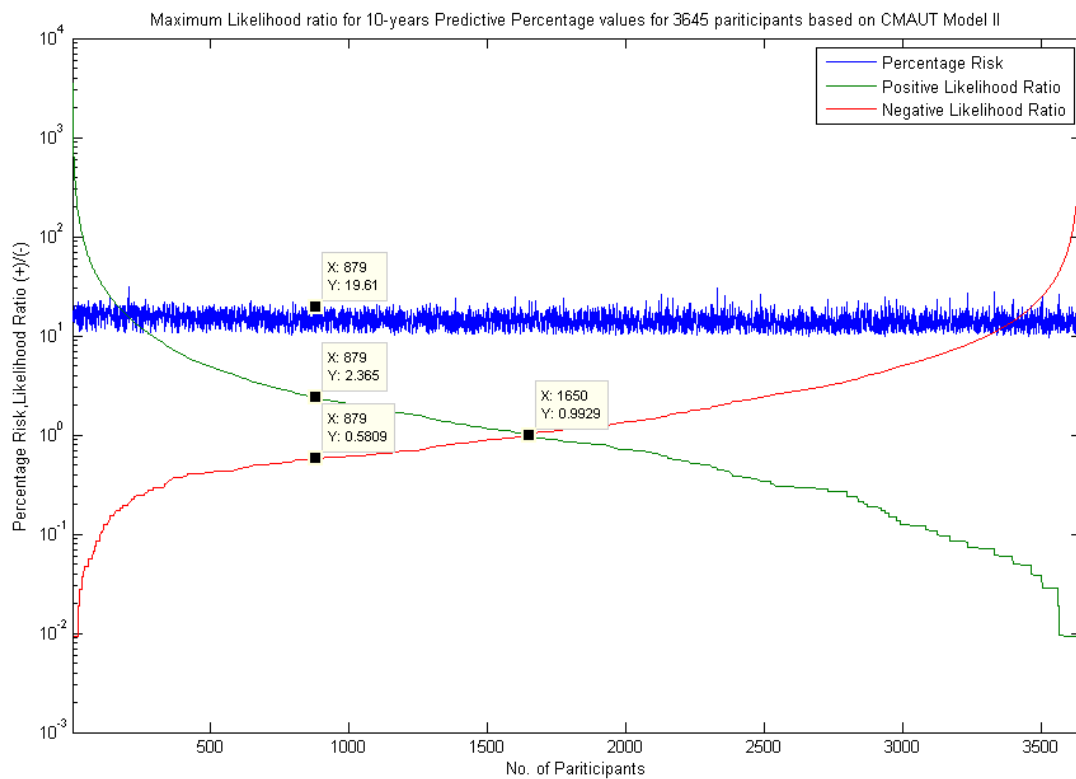


Figure 7.14: Likelihood Ratio curve of CMAUT Prognosis Risk Model – II

7.5 Summary:

In this chapter, two models of the CMAUT CVD prognosis framework were designed and built to prove the second part of the hypothesis. The second part of the hypothesis states that the framework can be used as an epidemiological tool to determine the percentage risk of a user been hypotensive in future. The model 1 was designed with the β_i coefficient values from the binary logistic regression analysis of the 4316 over 16 years old participants. The model 2 was designed using the β_i coefficient values from the 3645 participants, who are over 30 years old. The β_i coefficient values were used as weights for constructing the objective functions and statistical structured equations of models 1 and 2.

In both CMAUT CVD prognosis models 1 and 2, the Absolute Percentage Risk (u) was first computed for each participant. The predictive percentage risk $P(T)$ of each of the 3654 participants was computed as the arithmetical sum of the Absolute Percentage Risk (u) and risk factor $P(t)$ in 10 years' time. Based on these results the following verification metrics namely TPR, FPR, LRP and LRN were computed.

The resultant predictive percentage risk (PPR) values for each participant were used to plot the Receiver Operating Characteristics (ROC) to determine Area under the Curve (AUC) for models 1 and 2. Graphs for the sensitivity and selectivity of the participants were plotted against the NICE, (2006) of 20 % recommended criterion. Finally, the values of the positive and negative Likelihood ratio values were plotted against on the PIND of each participant. The predictive percentage risks from CMAUT models 1 and 2 as well as the graphs were presented in this Chapter 7 and also discussed in Chapter 9.

Chapter 8: Simulation of PPR with Web CHD Risk Calculators and Framingham Algorithms

8.0 Introduction

In this chapter, two CVD risk prediction techniques namely the Web-Based CHD Risk Calculators and Framingham Algorithms were analysed and discussed. First two Web-based CHD Risk Calculators, which were designed using Framingham Algorithm, were chosen. The two selected CHD Risk Calculators were used to determine the Predictive Percentage Risks (PPR) values of the selected 3645 participants discussed in Chapters 5 and 7. This was followed by discussing the three different types of Framingham Algorithms, which were used for building CVD Risk Prediction models (Wilson et al., 1998). The results of the PPR from the Framingham Algorithms and Web based CHD Risk Calculators were recorded and benchmarked against the two Prognosis CVD CMAUT framework models. The results are presented in this chapter and analysed in chapter 9.

8.1 Methodology used for the selection of the CHD web risk calculators

Internet based CHD risk calculators: E-health is the use of Internet based technological systems to manage and delivery health care. In this research, Internet technology refers to the usage of video, websites and scan images to capture clinical data that is used for the prognosis and diagnosis of diseases (Chuang et al, 2007). In computing, Internet is the physical internetworking infrastructure and the Web is the logical applications, which facilitate the exchange of data and information (Sheridan et al., 2003). In this research the terms Internet and Web will be used interchangeable to refer to the application software that resides on the Internet infrastructure. The PPRs obtained from the CMAUT frameworks are used to benchmark against the results from the two Web-Based CHD risk calculators.

The procedure used to shortlist and select the two CHD risk calculators for the benchmark exercise are:

1. Enter the keywords “CHD web risk calculators” into Google search engine (Sheridan et al., 2003). This gives a list of web based risk calculator and since the aim of this research is to compare the different types of Framingham equations. The search was narrowed to look for “CHD web risk calculators + Framingham Algorithm”.

2. The next step is to search for “UK web risk calculators + Framingham algorithm”. This advance search assisted in selecting only UK online risk calculators. The purpose for this selection is that the clinical data used for this research is from England hence the web calculator must be compatible with it. This process also reduced the search to 10 web calculators. See list of website in Appendix 8.
3. It was also identified that all the 10 web predictors were designed for different age groups starting from 32 to 94 years old. Again, they used different risk predictors and since the aim of this research is to predict the percentage risk using the participant’s clinical data, two web calculators with maximum clinical data were selected.

Finally the two CHD web risk calculators that were selected are:

- Internet model – I:- NHS BlackHeath centre (refer: <http://www.bhgp.co.uk/chdriskresult.asp>)
- Internet model – II:- Patient UK User Survey (refer: <http://www.patient.co.uk/doctor/Primary-Cardiovascular-Risk-Calculator.htm>).

The NHS BlackHeath Centre web site was selected because it has the NHS logo on the web page, which subsumes that it, was designed using the NICE guideline. This website also refers to Anderson et al, (1991) paper on CVD risk prediction that uses Framingham equation as the basis for the development of the CVD website.

Similarly, the Patient UK website was selected because it was designed and implemented using the Framingham algorithm (Anderson et al, 1991). Secondly, it was identified that the Patient UK website uses the main CVD risk factors required by prediction model for the computation of PPR based on the Framingham algorithm. The ten (10) websites from which these two Web-based calculators used for benchmarking were selected from are listed in Appendix 8.1.

- Determination of PPR using the Web-Based CVD Risk Calculators:

The data of each of the 3645 selected participants from the HSE, (2006) survey were input into the two web risk calculators and their 10 years Predictive Percentage Risk (PPR) values determined and recorded in the appropriate Tables.

The procedure for calculating the 10 years PPR of the selected 3645 participants using CHD web risk calculators is as follows: three participants from the HSE report were selected and used to illustrate the operation of the two selected websites. The same participants were used in Chapters 5 and 7 but their data have been reproduced in Table 8.0 for easy of reference.

Table 8.0 Data of participants used for illustration

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
42,831,101.00	No	No	35	Men	White	64.50	18.15	135.00	70.50	1.70	92.00	No	4.10	Yes	No	No

8.1.1 Determination of the 10 years PPR with Internet model 1:- NHS BlackHeath

The NHS BlackHeath website uses all the risk factors in Table 8.0 to compute the PPR values as indicated in the Framingham equations and explained in Anderson et al., (1991). The simulation exercise was conducted using the 3645 participants from the HSE, (2006) data set. Figure 8.0 below is the website in, which all the measurable and non-measurable risk factors for each participant were entered.

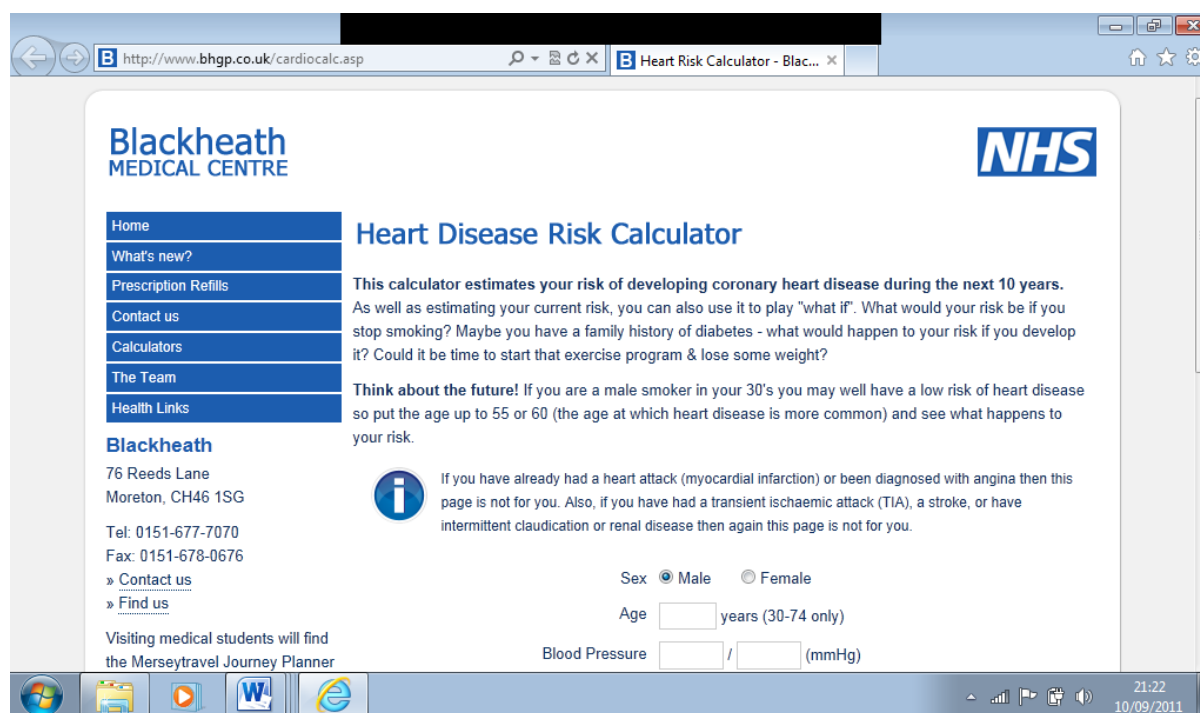


Figure 8.1: NHS BlackHeath Website for Heart Disease Risk calculator

The NHS BlackHeath web Calculator was designed for users, who are 30 to 74 years old. The measureable metrics used by the website are age, diastolic, systolic blood pressures, total cholesterol and HDL-Cholesterol. The non-measureable metrics are smoking and diabetic.

- The CVD data and Output results for the first participant:

The data of the first participant on the Table 8.0 was entered into the Heart Disease Risk Calculator and below in Figure 8.1 is the result. The CVD data used are PIND = 13,956,102.00, woman, white, Age = 60 years;; HB = 34.00 beat/sec, BMI = 13,71; BPH = 122.50 mgHH; BPL = 88.50 mgHH; HDL = 1.80 mmol; MAP = 100.00 mgHH; diabetic = No; TC = 5.20 mmol; Smoking = Yes; CVD = No; ECG = Yes; The output from the NHS BlackHeath Risk calculator is 7% in Figure 8.2:

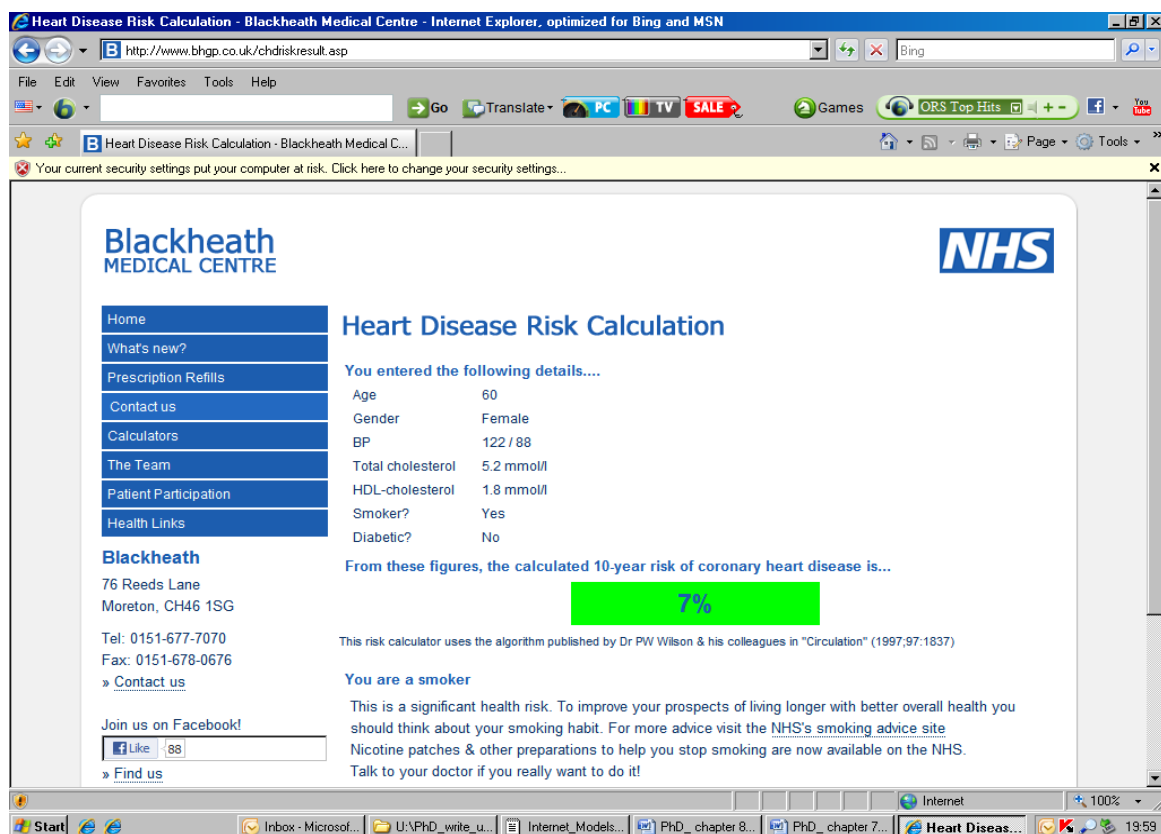


Figure 8.2: The output for participant one from the simulation NHS BlackHeath Website:

- The CVD data and Output for the second participant:

The data of the second participant on the Table 8.0 was entered into the Heart Disease Risk Calculator and below in Figure 8.1 is the result. The CVD data used are PIND = 63,535,102.00, woman, white, Age = 30 years; HB = 46.00 beat/sec, BMI = 13, 71; BPH = 120.00mgHH; BPL = 72.50 mgHH; HDL = 1.40 mmol; MAP = 89.50 mgHH; diabetics = No; TC = 4.50 mmol; Smoking = Yes; CVD = Yes; ECG = No; The output from the NHS BlackHeath Risk calculator is 1% in Figure 8.3:

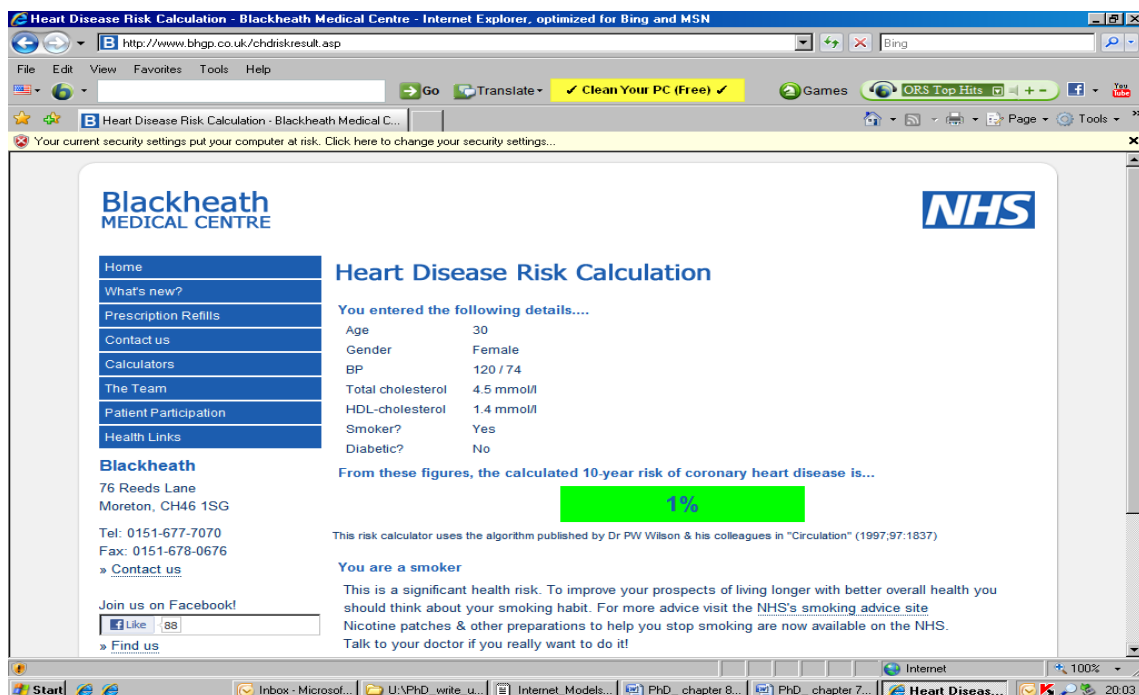


Figure 8.3: The output for participant two from the simulation NHS BlackHeath Website:

When this research was been conducted, the Heart Disease Risk Calculator available on the NHS BlackHeath Website was used for the simulation exercises. The results of the first 10 participants from the simulation exercises carried out using the then existing Risk Calculator are recorded in the Table 8.1 below in black colour and the rest of PPRs are in Appendix 8.

However, in the year 2012 and 2013 during the writing of this research, it was observed that the PPR results of the same participants have changed. The new Predictive Percentage Risk values are recorded in the Table 8.1 below in Red colour. This denotes that the NHS BlackHeath website have been updated to incorporate the new NICE, (2010) regulation although the website states that the copyright is for 2002 to 2011.

According to regulation in NICE, (2010), most CVD Risk Calculator websites that use Framingham algorithm give overestimated Predictive Percentage Risk values to users and this is discussed in Chapter 9. In this research the calculated PPR values recorded in the Table 8.1 below in black colour are used for the purpose of completeness and consistency. This is because before the introduction of NICE, (2010) regulation all the selected Web CVD Risk Calculators used the Framingham algorithm so the approach was the same and hence facilitates consistent and non-bias benchmarking.

- Simulation Results, Tables and Figures for the Web CVD calculators Model I:

The Tables and Figures in this section are the results of inputting the demographic and clinical data of each of the selected 3654 participants into the Web based CVD Risk calculator Internet Model I. The Table 8.1A is the raw data of the first 10 participants from the list of the 3654 selected participants, for the purpose of reference. Table 8.1B at the end of this Thesis contains the results of the first 30 participants and the results of the entire group are in the appendix Table 8.1C in electronic format. This is followed by Table 8.2A, which contains the results of the computed 10 years PPR values of the first 10 participants from Internet Model I. Table 8.2B, at the end of the Thesis contains the results of the first 10 participants and that of the entire group is in the Appendix Table 8.2C in electronic format.

The Table 8.3A shows the results of the computation of TPR, FPR, LRP and LRN for the NHS BlackHeath Website Internet Model I based on the PPR values for 10 years for the first 10 participants of the 3645 data set. Table 8.3B, at the end of this Thesis contains the results of the first 30 participants and that of the entire group are in the Appendix Table 8.3C in electronic format.

- Internet Model I Results of the 3645 participants in the category of over 30 years old:

Table 8.1A: Raw data of the first 10 participants

Pserial no.	Grp	BpI	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
71,831,101.00	No	No	66	Women	White	89.00	14.32	159.00	70.00	1.90	99.50	No	6.90	No	No	Yes
34,031,101.00	No	No	84	Women	White	48.50	16.17	112.00	63.50	2.20	80.00	No	5.00	Yes	Yes	Yes
72,604,102.00	No	No	59	Women	White	36.50	16.19	109.50	73.00	2.00	85.00	No	6.00	No	No	No
13,008,101.00	Yes	Yes	50	Women	White	43.00	16.65	117.00	74.00	1.70	88.50	No	6.00	Yes	Yes	Yes
39,139,101.00	No	No	34	Women	White	48.00	16.81	102.00	54.00	1.80	70.00	No	6.50	Yes	No	No
47,856,102.00	No	No	51	Women	White	44.00	16.85	100.50	56.50	1.90	71.00	No	5.10	No	No	No
37,710,101.00	No	No	61	Women	White	43.00	17.43	120.00	77.00	1.20	91.50	No	5.50	No	No	No
54,256,101.00	No	No	31	Women	White	40.00	17.72	124.00	84.00	2.00	97.00	No	3.90	Yes	No	No

Table 8.2A: Predicative Percentage Risks for 10 years of the first 10 participants based on Internet

Model – I NHS BlackHeath centre (ref: <http://www.bhgp.co.uk/chdriskresult.asp>).

Pserial no.	Age	Sex	BMI	BPH	BPL	HDL	DIA	TC	SMK	ECG	% PR
13,956,102.00	60	Women	13.20	122.50	88.50	1.80	No	5.20	Yes	Yes	7 7
63,535,102.00	30	Women	13.71	120.00	74.00	1.40	No	4.50	Yes	No	1 1
71,831,101.00	66	Women	14.32	159.00	70.00	1.90	No	6.90	No	Yes	8 8
34,031,101.00	84	Women	16.17	112.00	63.50	2.20	No	5.00	Yes	Yes	3 NA
72,604,102.00	59	Women	16.19	109.50	73.00	2.00	No	6.00	No	No	3 3
13,008,101.00	50	Women	16.65	117.00	74.00	1.70	No	6.00	Yes	Yes	1 3
39,139,101.00	34	Women	16.81	102.00	54.00	1.80	No	6.50	Yes	No	2 1
47,856,102.00	51	Women	16.85	100.50	56.50	1.90	No	5.10	No	No	11 2
37,710,101.00	61	Women	17.43	120.00	77.00	1.20	No	5.50	No	No	0 11
54,256,101.00	31	Women	17.72	124.00	84.00	2.00	No	3.90	Yes	No	1 0

Table 8.3A: Calculation of TPR, FPR, LRP, and LRN, for the Internet Model I for the first 10

Participants NHS BlackHeath centre (ref: <http://www.bhgp.co.uk/chdriskresult.asp>)

Pserial no.	Grp	Bpl	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRM
13,956,102.00	No	No	60	Women	7	0	1	1	0.99968	3125	0
63,535,102.00	Yes	Yes	30	Women	1	0	1	1	0.99936	1562.5	0
71,831,101.00	No	No	66	Women	8	0	1	1	0.999041	1042.753	0
34,031,101.00	No	No	84	Women	3	0	1	1	0.998721	781.8608	0
72,604,102.00	No	No	59	Women	3	0	1	1	0.998401	625.3909	0
13,008,101.00	Yes	Yes	50	Women	1	0	1	1	0.998081	521.1047	0
39,139,101.00	No	No	34	Women	2	0	1	1	0.997761	446.628	0
47,856,102.00	No	No	51	Women	11	0	1	1	0.997442	390.9304	0
37,710,101.00	No	No	61	Women	0	0	1	1	0.997122	347.4635	0
54,256,101.00	No	No	31	Women	1	0	1	1	0.996802	312.6954	0

Figure 8.1 shows the graph of the computed 10 years PPR value of each of the 3645 participants plotted against their individual PIND. The data used to plot the graph is from Table 8.2C. Figure 8.2 is the prediction accuracy graph of the computed results of TPR against the FPR of each of the 3645 participants. The data used to plot the AUC graph is from Table 8.3C. The AUC area for Internet Model I was calculated, by summing the PPR data points using the trapezoidal method. The actual AUC is obtained by subtracting the sum of the entire PPR data points from the sum of all the diagonal reference data points as discussed in chapter 3.

Figure 8.3 shows the discriminatory accuracy or discriminatory ability graph for the Internet Model I. This was constructed by plotting the sensitivity and selectivity values of each of the 3645 participants in Table 8.3C. The graph of the sensitivity and selectivity values against the NICE, (2006) recommended criterion of 20% and the interception are discussed in chapter 9. Table 8.3C contains the results of the computed positive and the negative Likelihood ratio of the entire 3645 participants.

Figure 8.4 is the performance accuracy graph of the PPR values for each participant's results from the Internet Model I. This is based on plotting all the values of the positive and negative Likelihood ratios on the Y-axis and the PIND for each participant on the X-axis. The performance accuracy graph is discussed in chapter 9.

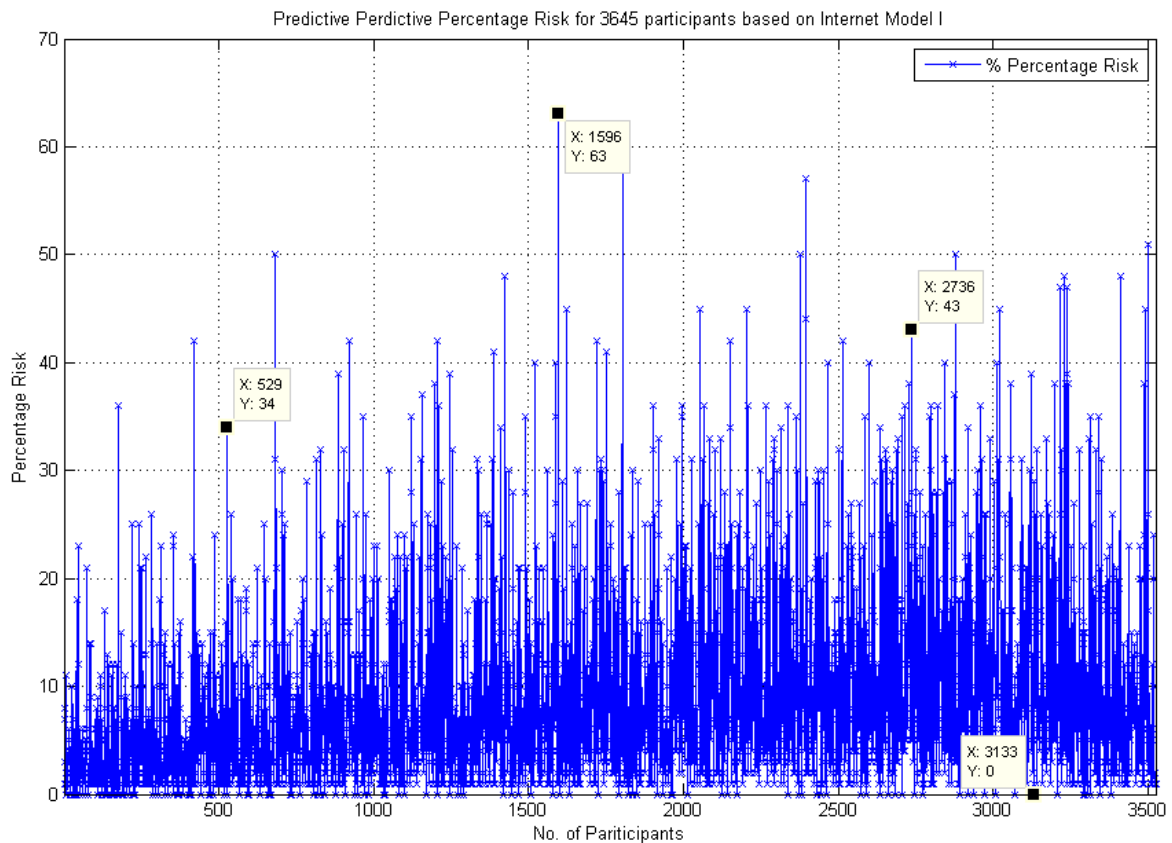


Figure 8.1: Predictive Percentage Risk for 3645 Participants based on Internet Model I

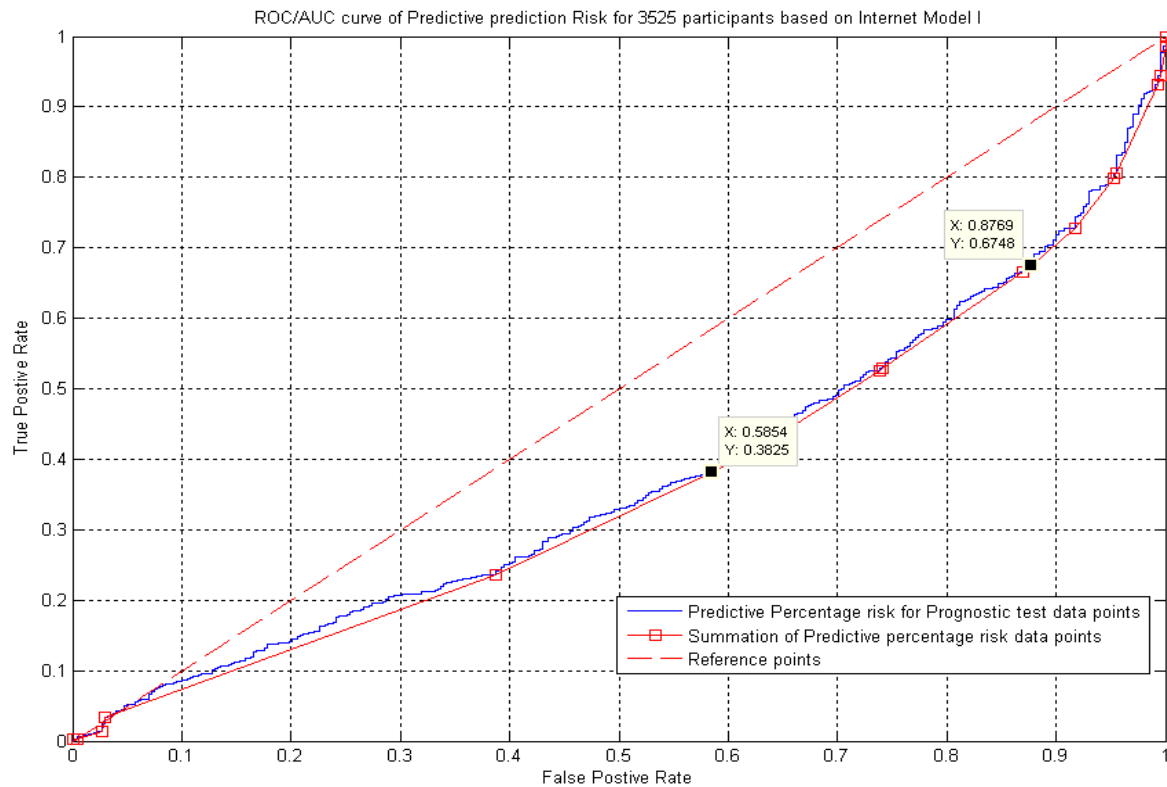


Figure 8.2: ROC/AUC for the Internet Model I

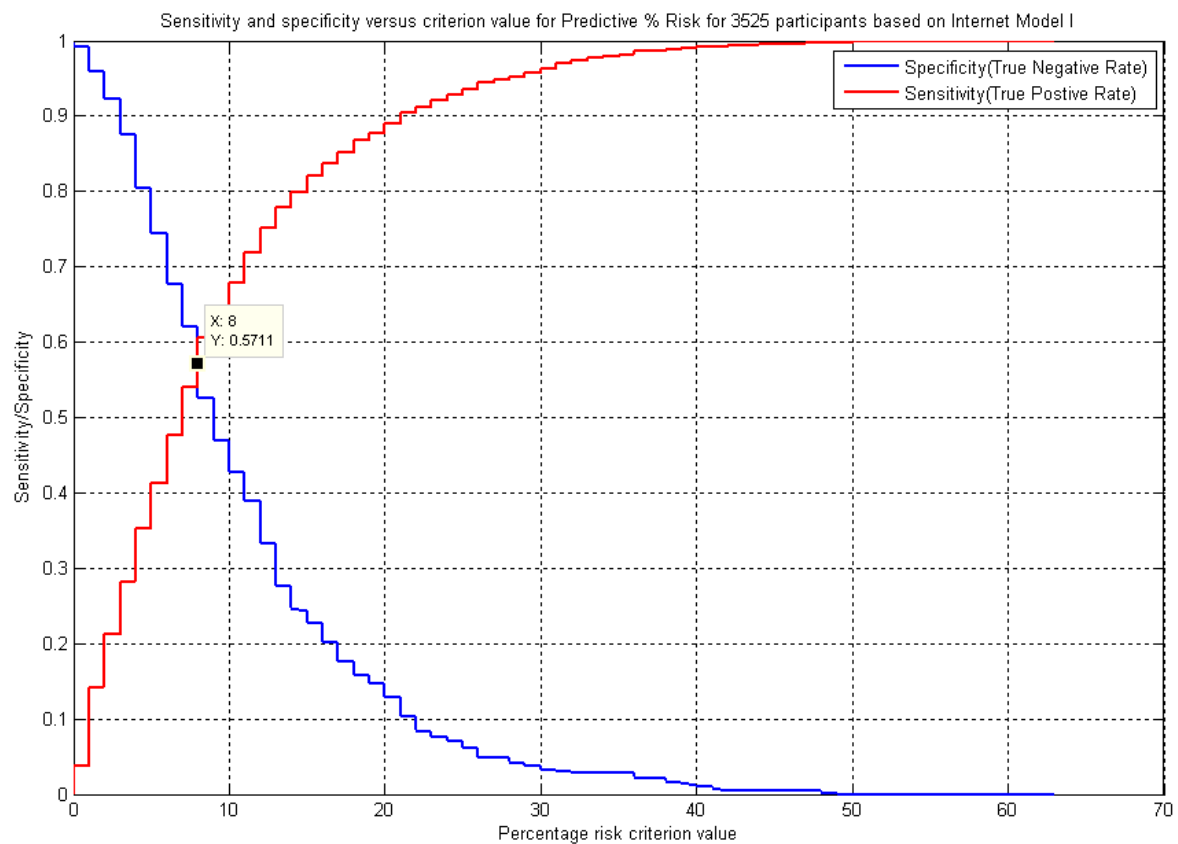


Figure 8.3: Sensitivity and specificity of the Internet model I

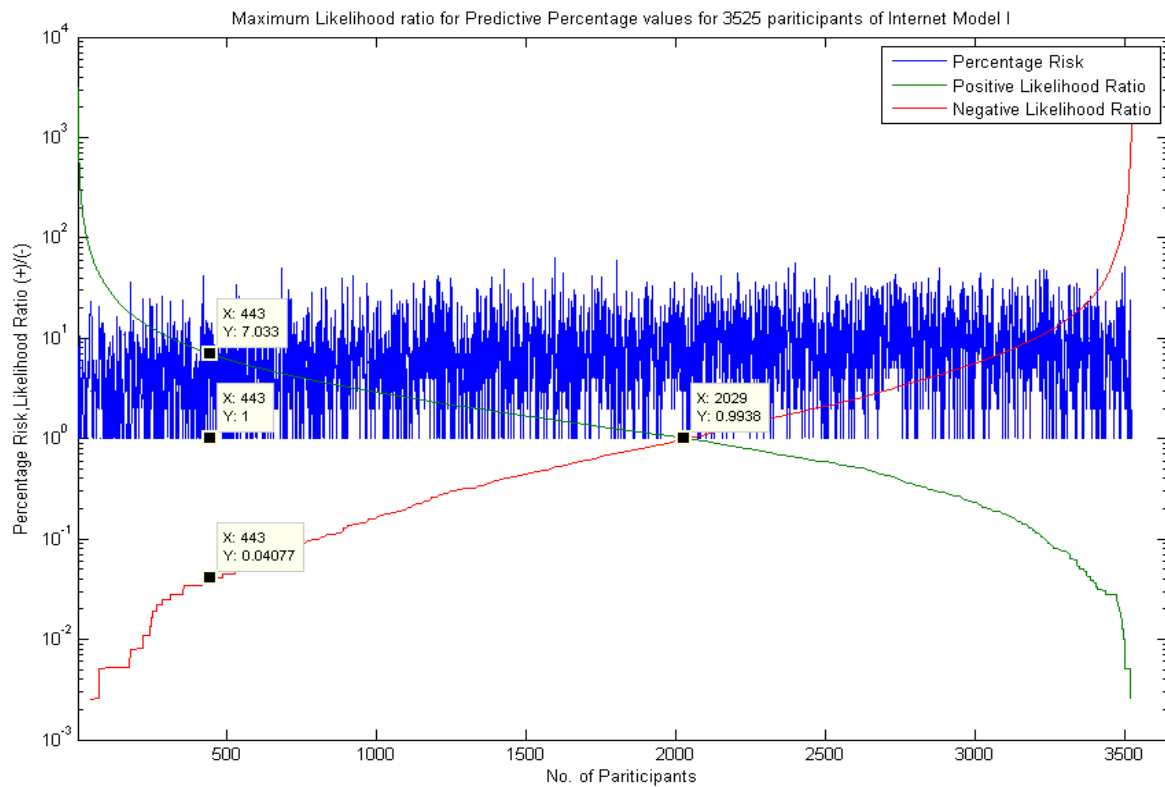


Figure 8.4: Maximum likelihood ratio for the Internet model I

8.1.2 Simulation of the 10 years PPR with Internet model 2:- Patient UK

The Patient.co.uk website was designed with the Framingham risk equation by Anderson et al., (1991). In addition, the website uses the adjustment proposed by the Joint British Societies and all the risk factors in Table 8.1 to compute the PPR values. The simulation exercise was conducted by using 3465 participants from the HSE, (2006) data set. Figure 8.5 is the website in which all the measurable and non-measurable risk factors for each participant were entered.

- Determination of the 10 years PPR with Internet model 2 Patient.co.uk:

Again, each of the selected 3645 participants from the HSE, (2006) report were input into the Patient.co.uk, web risk calculator and their 10 years Predicted Percentage Risk computed and recorded in Table 8.5. For the purpose of illustration, two simulation exercises were conducted using data set of two participants from the HSE, (2006). Figure 8.5 is the website in which all the measurable and non-measurable risk factors for each of the participant were entered. The website is notated as Internet model-II Patient.co.uk: (from <http://www.patient.co.uk/doctor/Primary-Cardiovascular-Risk-Calculator.htm>).

Primary Cardiovascular Risk Calculator

This calculator uses the Framingham risk equation¹ and the adjustments as suggested by the Joint British Societies' (JBS2) paper² and the JBS Cardiovascular Risk Assessor.³

The latest National Institute for Health and Clinical Excellence (NICE) Guidance (2010) does not recommend any particular risk calculator, but does emphasise that any which are based on the Framingham risk equation may overestimate risk in UK populations.⁴ An alternative is the Qrisk@2 calculator [here](#).

Cardiovascular Risk Calculator For Primary Prevention			
This calculator should not be used if patient has known CVD or diabetes (already known to be at high risk)			
Age (30-74)	<input type="text"/>	Smoking Status	<input type="text" value="Non Smoker"/>
Sex	<input type="text" value="Male"/>	Glucose	<input type="text" value="Normal"/>
Systolic BP	<input type="text"/>	LVH	<input type="text" value="No LVH"/>
Diastolic BP	<input type="text"/>	Central Obesity	<input type="text" value="No"/>
Total Cholesterol	<input type="text"/>	South Asian Origin	<input type="text" value="No"/>
HDL Cholesterol	<input type="text"/>	Family History of CVD (Men <55 and women <65 years)	<input type="text" value="No FH"/>

Lower Your Cholesterol
Find the easy way to lower your cholesterol with Flora pro.activ
www.floraproactiv.co.uk

Statin Side Effects
Get The Real Truth
Doctors Won't Tell You
About Cholesterol
Drugs
TheCholesterolTruth.com/St...

Figure 8.5: Patient UK Website for Primary Cardiovascular Risk calculator

The Patient.co.uk Web CVD Risk Calculator was designed for users, who are between the ages 30 and 74 years old. The measureable metrics used by the site are age, systolic, diastolic blood pressures, BMI, total cholesterol and HDL-Cholesterol. The non-measurable metrics are smoking, diabetic, ethnicity, CVD and ECG.

- The CVD data and Output for the first participant:

The data for the first participant in Table 8.1 were entered into the Cardiovascular Risk Calculator and the result is shown in Figure 8.6. The CVD data used were PIND = 13,956,102.00, woman, white, Age = 60 years;; HB = 34.00 beat/sec, BMI = 13,71; BPH = 122.50 mgHH; BPL = 88.50 mgHH; HDL = 1.80 mmol; MAP = 100.00; diabetic = No; TC = 5.20 mmol; Smoking = Yes; CVD = No; ECG = Yes; The output value of the Systolic Blood Pressure from Patient.co.uk Web CVD Risk Calculator is 21% as shown in Figure 8.6:

Cardiovascular Risk Calculator For Primary Prevention

This calculator should not be used if patient has known CVD or diabetes (already known to be at high risk)

Age (30-74): 60 Smoking Status: Smoker

Sex: Female Glucose: Normal

Systolic BP: 122.50 LVH: LVH on ECG

Diastolic BP: 88.50 Central Obesity: No

Total Cholesterol: 5.20 South Asian Origin: No

HDL Cholesterol: 1.80 Family History of CVD (Men): No FH

Total /HDL Ratio: 2.89

Calculate Clear Fields

Serum TG mmol/L:

Using Systolic BP prediction, the 10-year risk of JBS CVD Risk is 21 %

The equivalent risk calculation with diastolic BP is 33 %

The following patients will have higher risks:

Figure 8.6: The output for participant one from the simulation Patient UK Website:

- The CVD data and Output for the second participant:

The data for the second participant on the Table 8.1 were entered into the Heart Disease Risk Calculator and the result is shown in Figure 8.6. The CVD data used is PIND = 63,535,102.00, woman, white, Age = 30 years. The other parameters are HB = 46.00 beat/sec, BMI = 13, 71; BPH = 120.00mgHH; BPL = 72.50 mgHH; HDL = 1.40 mmol; MAP = 89.50 mgHH; diabetics = No; TC = 4.50 mmol; Smoking = Yes; CVD = Yes; ECG = No; The output PPR value for the Systolic Blood Pressure from the Patient.co.uk Web CVD Risk Calculator is 0.5 % as shown in Figure 8.7:

Cardiovascular Risk Calculator For Primary Prevention

This calculator should not be used if patient has known CVD or diabetes (already known to be at high risk)

Age (30-74): 30 Smoking Status: Smoker
 Sex: Female Glucose: Normal
 Systolic BP: 120 LVH: No LVH
 Diastolic BP: 74 Central Obesity: No
 Total Cholesterol: 4.5 South Asian Origin: No
 HDL Cholesterol: 1.4 Family History of CVD (Men): Significant FH CVD
 Total /HDL Ratio: 3.21

Calculate Clear Fields

Serum TG mmol/L:

Using Systolic BP prediction, the 10-year risk of JBS CVD Risk is: <0.5 %

The equivalent risk calculation with diastolic BP is: <0.5 %

The following patients will have higher risks:

- Significant family history (men <55 and women <65 years with one first-degree relative).

Latest posts:

- Patient care across health communities - a how-to checklist for CCGs
- Decisions, decisions
- Patient awareness
- The right-to-die - a test for overall freedom?
- Lesions disseminated in time and space - MS awareness week

Read more »

Advertisements:

- Private Medical Insurance for as little as £1 a Day
- Get Priority Access to Consultants
- Avoid NHS Queues

Figure 8.7: The output for participant two from the Patient UK CVD Risk calculator:

When this research was been carried out in the year 2010, the old Web CVD Risk Calculator version 24 of 2009 was available. Therefore the simulation exercises were conducted using the then existing website and the PPR results are shown in the Tables below and the rest of the results are in Appendix 8. The results of the first 10 participants from the simulation exercises carried out on the Risk Calculator version 24 are recorded in Table 8.4A below in black colour and the rest are in Appendix 8.

However, in 2011 and 2012, the www.patient.co.uk website was updated and the Qrisk@2 risk calculator was incorporated into the website in accordance with the NICE, (2010) recommendations. For the purpose of completeness, the Table 8.4A below has the calculated PPR values from the current website Qrisk@2, which are entered into the Table 8.4A using red colour. However, these results are not used for analysis because the technique used is different from the Framingham algorithm discussed in this research.

- Version 24 (2009) and Version 28 (2011) of the www.patient.co.uk website:

In the year 2009 www.patient.co.uk launched a CVD web risk calculator version 24 that uses the Framingham algorithm. This site was used to determine the 10 years predictive percentage risk values of the participants in the HSE, (2006) report been hypertensive and the results were recorded in Table 8.4 below.

However, in the year 2011, www.patient.co.uk updated their web risk calculator because according to NICE, (2010) using the Framingham risk algorithm gives an overestimated percentage risk values for the UK populations. This is stated in the NICE Clinical Guideline (May, 2008) that was amended in May 2010 therefore www.patient.co.uk now uses Qrisk®2 as an alternative CVD risk predictor. Other issues regarding the different between the various CVD risk calculator websites and results of the simulation exercises are discussed further in chapter 9.

The Table 8.5A shows the results of the computation of TPR, FPR, LRP and LRN for the Patient UK Website Internet Model II based on the PPR for 10 years for the first 10 participants of the 3645 data set. The Table 8.5B, at the end of this Thesis contains the results of the first 30 participants and the result of the entire group is in the Appendix Table 8.5C in electronic format.

Table 8.4: Predicative Percentage Risks for 10 years for the first 10 participants based on Internet

Model – II Patient UK User Survey (ref: <http://www.patient.co.uk/doctor/Primary-Cardiovascular-Risk-Calculator.htm>)

Pserial no.	Age	Sex	BMI	BPH	BPL	HDL	DIA	TC	SMK	ECG	% PR
13,956,102.00	60	Women	13.20	122.50	88.50	1.80	No	5.20	Yes	Yes	9 21
63,535,102.00	30	Women	13.71	120.00	74.00	1.40	No	4.50	Yes	No	4.3 0.5
71,831,101.00	66	Women	14.32	159.00	70.00	1.90	No	6.90	No	Yes	20.9 30
34,031,101.00	84	Women	16.17	112.00	63.50	2.20	No	5.00	Yes	Yes	17.5 NA
72,604,102.00	59	Women	16.19	109.50	73.00	2.00	No	6.00	No	No	11.8 4
13,008,101.00	50	Women	16.65	117.00	74.00	1.70	No	6.00	Yes	Yes	9.3 17*
39,139,101.00	34	Women	16.81	102.00	54.00	1.80	No	6.50	Yes	No	8.7 0.5
47,856,102.00	51	Women	16.85	100.50	56.50	1.90	No	5.10	No	No	10 2
37,710,101.00	61	Women	17.43	120.00	77.00	1.20	No	5.50	No	No	7.5 8
54,256,101.00	31	Women	17.72	124.00	84.00	2.00	No	3.90	Yes	No	4.8 0.5

Table 8.5: Calculation of TPR, FPR, LRP, and LRN, for the Internet Model – II for the first 10 participants Patient UK User Survey (ref: <http://www.patient.co.uk/doctor/Primary-Cardiovascular-Risk-Calculator.htm>)

Pserial no.	Grp	Bpl	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRN
13,956,102.00	No	No	60	Women	9	0	1	1	0.9997	3584.22	0
63,535,102.00	Yes	Yes	30	Women	4.3	0	1	1	0.9994	1788.90	0
71,831,101.00	No	No	66	Women	20.9	1	0	0.984848	0.9994	1761.80	0.01516
34,031,101.00	No	No	84	Women	17.5	0	1	0.984848	0.9991	1175.23	0.01516
72,604,102.00	No	No	59	Women	11.8	0	1	0.984848	0.9988	880.90	0.01516
13,008,101.00	Yes	Yes	50	Women	9.3	0	1	0.984848	0.9986	704.97	0.01517
39,139,101.00	No	No	34	Women	8.7	0	1	0.984848	0.9983	587.61	0.01517
47,856,102.00	No	No	51	Women	10	0	1	0.984848	0.9980	503.50	0.01518
37,710,101.00	No	No	61	Women	7.5	0	1	0.984848	0.9977	440.64	0.01518
54,256,101.00	No	No	31	Women	4.8	0	1	0.984848	0.9974	391.58	0.01519

Figure 8.8 shows the graph of the computed 10 years PPR value of each of the 3645 participants against their individual PIND. The data used for plotting the graph is from Table 8.4C. Figure 8.9 is the prediction accuracy graph of the computed values of TPR against the FPR value of each of the 3645 participants. This data used to plot the AUC graph is from Table 8.4C and the AUC area for Internet Model 2 was calculated, by summing all the PPR data points using the trapezoidal method. The area is obtained by subtracting the sum of all the PPR data points from the sum of all the diagonal reference data points as discussed in chapter 3.

Figure 8.10 is the discriminatory accuracy or discriminatory ability graph of the Internet Model II. This was constructed by plotting the sensitivity and selectivity of each of the 3645 participants in Table 8.5C. The graph of the sensitivity and selectivity values against the NICE, (2006) recommended 20% criterion and the interception of the graph are discussed in chapter 9. Table 8.5C contains the computation results of the positive and the negative Likelihood ratio values for all the 3645 participants. The Figure 8.11 is the performance accuracy graph of the PPR values of each participant's value from the Internet Model II. This is obtained by plotting all the values of the positive and negative Likelihood ratios on the Y-axis and the PIND for each participant on the X-axis. The graph is discussed in chapter 9.

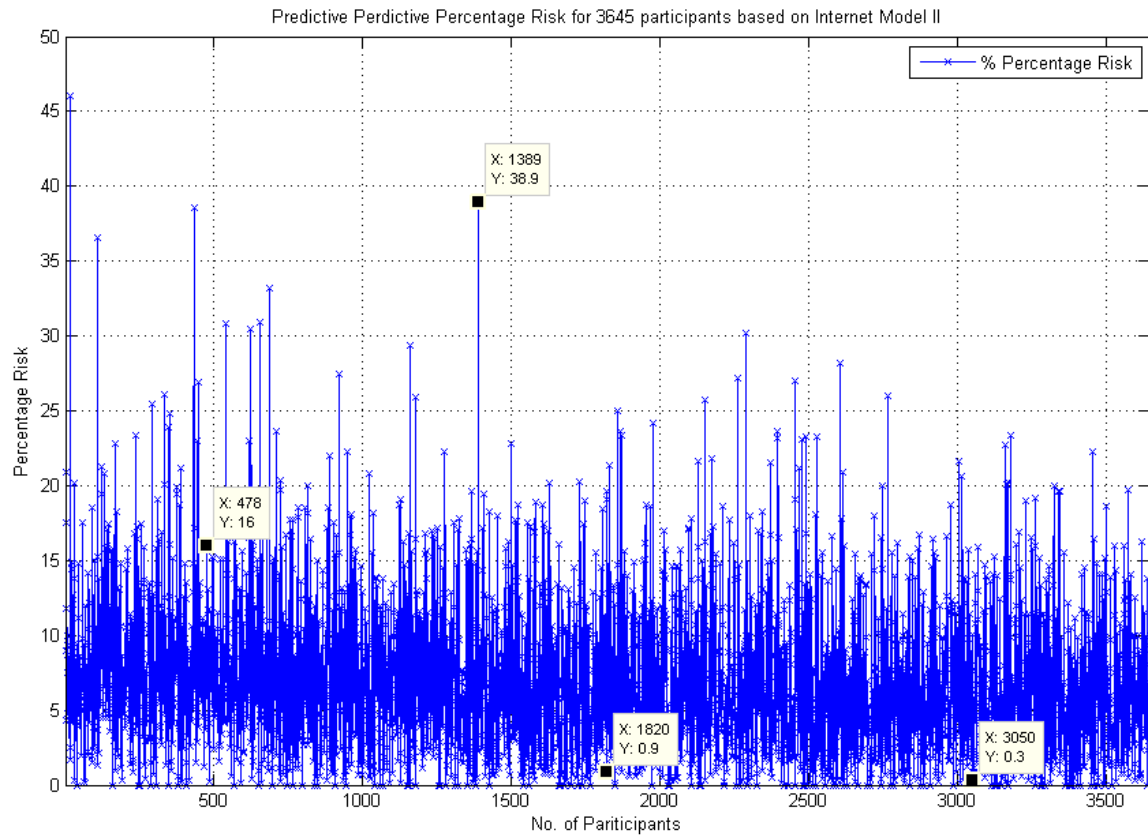


Figure 8.8: Predictive Percentage Risk for 3645 participants for Internet Model II

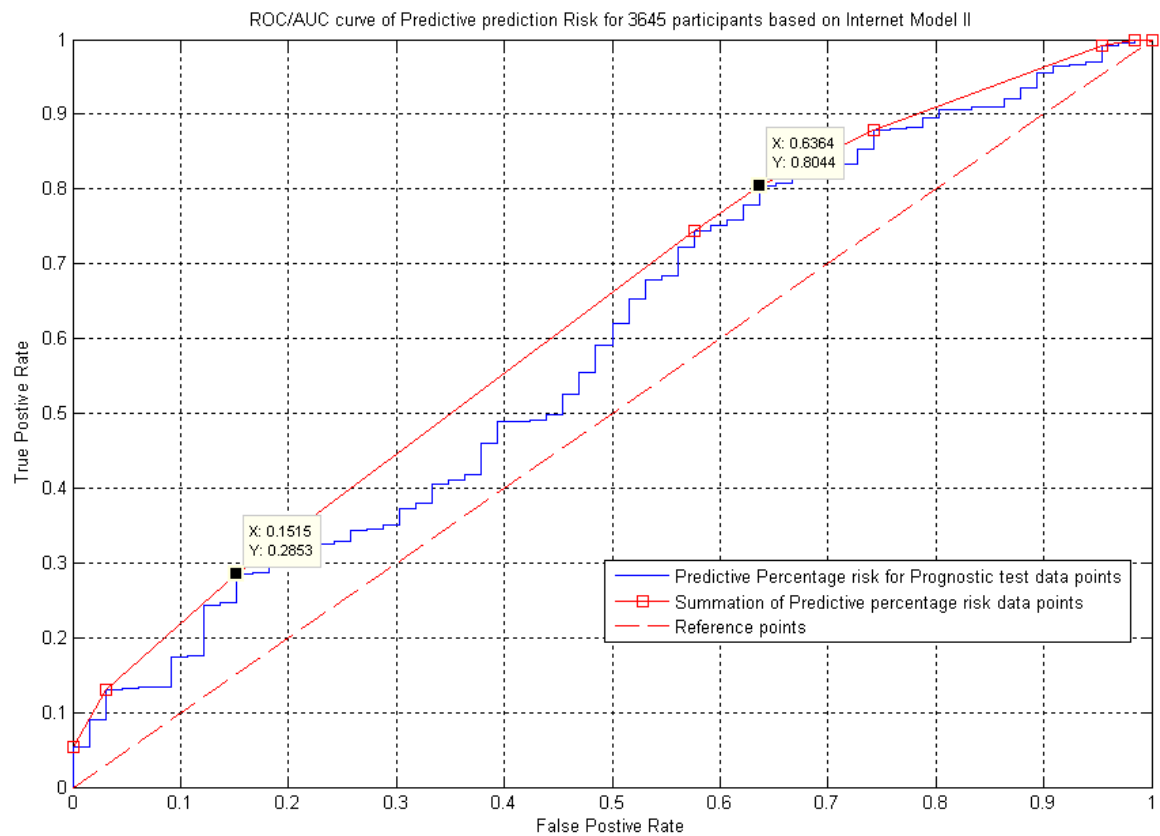


Figure 8.9: ROC/AUC of predictive prediction risk for 3645 for Internet Model II

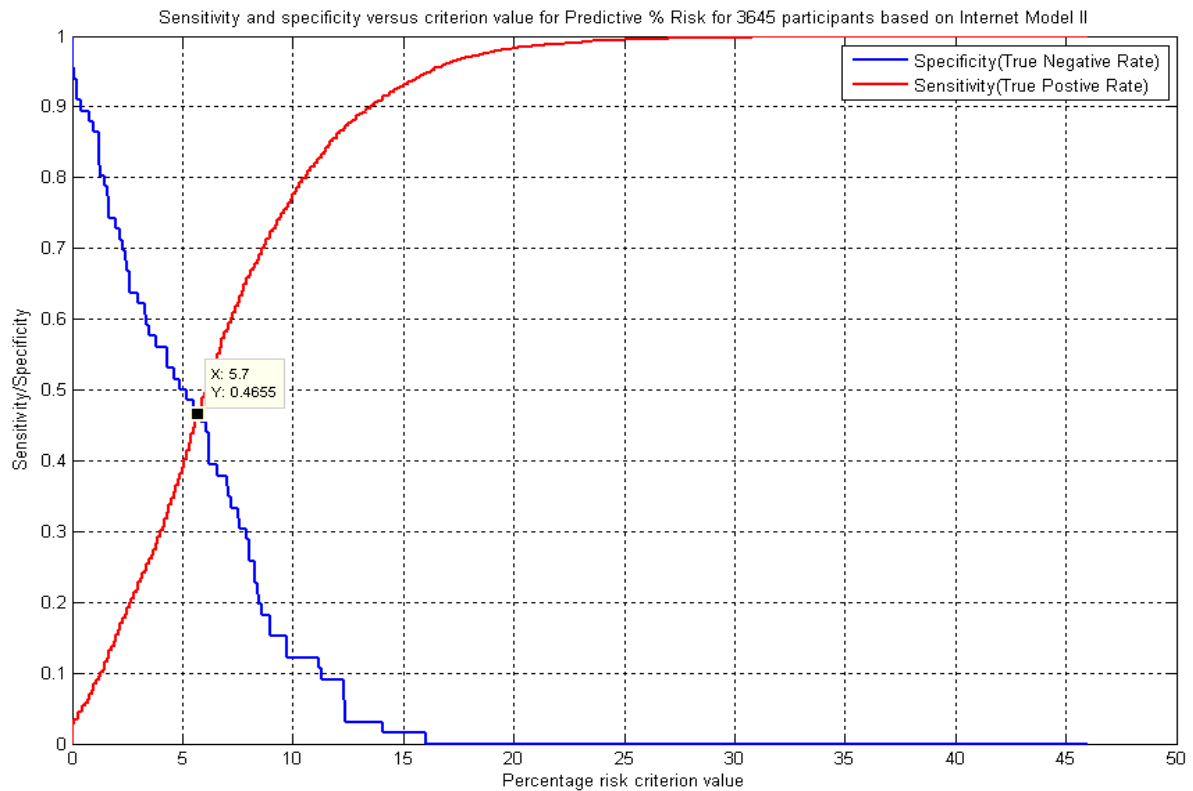


Figure 8.10: Sensitivity and specificity for Internet Model II

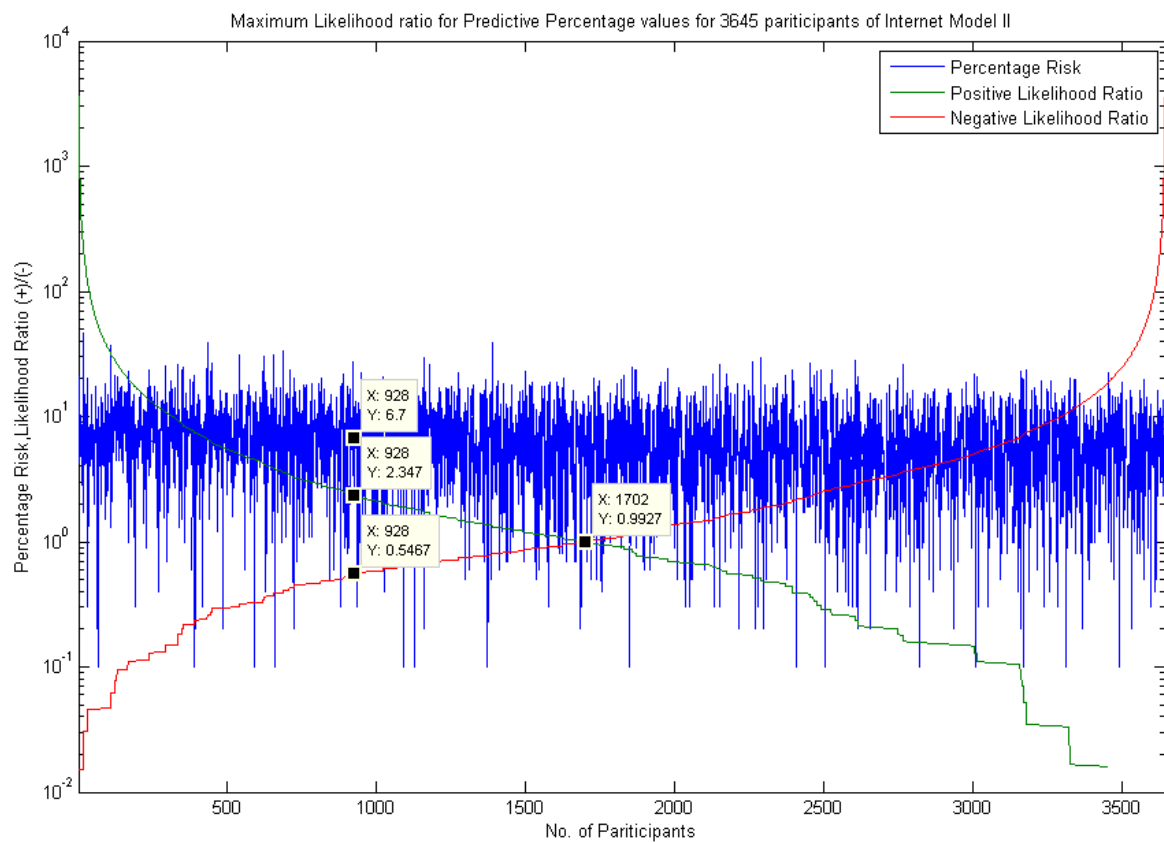


Figure 8.11: Maximum Likelihood ratio for Internet model II

8.2 Framingham Algorithm for the determination of PPR values

The Framingham equation is the second method, which output PPR values were used to benchmark the PPR results from the CMAUT Prognosis framework. In chapter 7, the Framingham equation was introduced and it was explained that this CVD prediction technique is used for calculating and predicting the probability percentage risk of a person having CVD disease over a periodic of time (Anderson et al, 1991). The Framingham algorithm uses the CVD risk factors both clinical measurable attributes such as BPH, BPL, TC, HDL, BMI, ECG and the non- measurable attributes risk such as the behaviour of the person which are smoking, origin, age and sex to predict the CVD percentage risk (Sheridan et al., 2003).

Framingham algorithm was developed using the regression statistical models. The algorithm was validated by conducting a series of follow-ups to confirm the estimated percentage risk values after the specified time frame of 4, 10 and 12 years (D'Agostino et al., 2001). In this research, three versions of the Framingham algorithm are evaluated and their outcomes compared with the CMAUT CVD frameworks.

They are:

1. The original Framingham algorithm from Anderson group (Anderson et al., 1991)
2. The England version of Framingham algorithm (Brindle et al., 2003)
3. The International Task Force version of Framingham algorithm (Zgibor et al., 2006);

8.2.1 Original Framingham algorithm from USA

The original document from K.M. Anderson group (Anderson et al., 1991), which is known as “An updated coronary risk profile; A statement for health professionals” has a step by step calculation on the procedure for the determination of CVD Percentage Predictive Risk (PPR). In this research, Systolic Blood Pressure (SBP) is used; therefore the equations discussed below incorporate the SBP parameters. Again, after each calculation, the components used in the equations are explained. The constants and variables used in these original equations are based on the survey conducted by the K.M. Anderson group in USA. They also used the statistical regression modelling and follow-up techniques to confirm their findings.

- Systolic Blood Pressure Equation

There are differences in the formulae used for the computation of Predictive Percentage Risk (PPR) for men and women; however both calculations start with the same approach.

Step 1: is the computation of the initial value **a**, which is based on the values of the risk factors measurements stated in Anderson et al. (1999).

$$\begin{aligned} \mathbf{a} = & 11.1122 - 0.9119 \times \log(\text{SBP}) - 0.2767 \times \text{Smoking} - 0.7181 \times \log(\text{cholesterol/HDL}) - \\ & 0.5865 \times \text{ECG-LVH} \end{aligned} \quad (8.1)$$

Step 2: is the computation of the second interim value **m**, this uses the participant's age and diabetes condition but the equation is different for men and women.

For men, compute

$$\mathbf{m} = \mathbf{a} - 1.4792 \times \log(\text{age}) - 0.1759 \times \text{diabetes} \quad (8.2a)$$

For women, compute

$$\mathbf{m} = \mathbf{a} - 5.8549 + 1.8515 \times [\log(\text{age}/74)] - 0.3758 \times \text{diabetes} \quad (8.2b)$$

Step 3: is the calculation of the μ and σ for both sexes, which are computed as follows:

$$\mu = 4.4181 + m \quad (8.3)$$

$$\sigma = \exp(-0.3155 - 0.2784 \times m) \quad (8.4)$$

Step 4: is the selection of the number of years (**t**) required for the risk prediction which is between 4 to 12 years. The selected (**t**) is used to calculate the predictive time factor (**u**) using the computed values of the location (μ) and scale (σ) parameters. In this work, **t** = 10 years.

$$\mathbf{u} = (\log(t) - \mu) / \sigma. \quad (8.5)$$

The predicted risk probability for the time **t** is

$$\mathbf{p} = 1 - \exp(-e^{\mathbf{u}}) \quad (8.6)$$

In percentage the PPR = (**p** x 100)%.

- Case study for the illustration of the determination of PPR Risk value

In this case study, two participants a male and female both 55-year-old were selected. For the purpose of illustration both have the following measurable and non-measurable CVD risk factor values.

Their SBP is 130 mm Hg, total cholesterol of 240 mg/dl, HDL cholesterol of 45 mg/dl and both smoke cigarettes. Again, their data indicates that they are neither diabetic nor have been diagnosed with ECG-LVH.

In order to compute their PPR in 10 years, the non-measurable risk factors namely smoking, electrocardiographic left ventricular hypertrophy (ECG-LVH), and diabetes are set to 1 when present and 0 when absent. Whereas Systolic Blood Pressure is measured in mm Hg, total cholesterol in mg/dl, HDL cholesterol in mg/dl and the age is in years.

- Below is the procedure for calculating the PPR in 10 years for the two participants:

First, we compute $a = 11.1122 - 0.9119 \times \log(130) - 0.2767(1) - 0.7181 \times \log(240/45) = 5.1947$

For a man, the computation:

$$m = 5.1947 - 1.4792 \times \log(55) - 0.1759 \times (0) = -0.7329$$

$$\mu = -0.7329 + 4.4181 = 3.685$$

$$\sigma = \exp[-0.3155 - 0.2784 \times (-0.7329)] = 0.894$$

For a woman, the computation is

$$m = 5.1947 - 5.8549 + 1.8515 \times [\log(55/74)]^2 - 0.3758(0) = -0.4972$$

$$\mu = -0.4972 + 4.4181 = 3.921$$

$$\sigma = \exp[-0.3155 - 0.2784 \times (-0.4972)] = 0.8377$$

As stated in the case study the prediction time t is 10 years, and the computation for men is:

$$u = (\log(10) - 3.685) / 0.894 = 1.546$$

$$p = 1 - \exp(-e^{-1.546}) = 0.192$$

For a woman, the computation is

$$u = (\log(10) - 3.921) / 0.8377 = 1.932$$

$$p = 1 - \exp(-e^{-1.932}) = 0.135$$

Therefore from the equation the Percentage Predictive Risk (PPR) = $(p \times 100) \%$

For the man the p value is 0.192 and therefore the PPR is 19.2%.

For the woman the $p = 0.135$ and therefore the PPR is 13.5 %.

8.2.2 Framingham algorithm from British Perspective:

In British, the Framingham Equation was modified and applied by Brindle et al., (2003) in their research work; <http://www.bmj.com/cgi/reprint/327/7426/1267>. Below is a brief discussion on how Brindle et al., (2003) adapted the Framingham equation in Anderson et al., (1999) to analyse CHD among British men. In their research, Brindle et al., (2003) modified some of the constants and variables in the original Framingham equations to match their findings in England.

In their equations just as in the original Framingham algorithms, the Systolic blood pressure (SBP) values are measured in mmHg, age in years, total cholesterol and high density lipoprotein cholesterol in mmol. Whereas the non-measurable variables such as smoking, electrocardiographic left ventricular hypertrophy, and diabetes are set to 1 when present and 0 when absent.

- For men for England, UK

Below is the procedure used for the computation of the Percentage Predictive Risk of British male having CHD within a specified time. The step wise procedure applied by the Framingham risk equations for the determination of coronary heart disease death (B1) and coronary heart disease events (B2) in men over 10 years are:-

- Framingham risk equations for CHD events (B2) of men within 10 years:

Step 1: involves the computation of either the coronary heart disease death (B1) or the coronary heart disease events (B2); Both equations, unlike the original Framingham equations discussed above combines the measurable and non- measurable attribute values together to form one long equation known as μ . This research focuses on Coronary Heart Disease events (CHD) and not the coronary heart death, therefore equation (B2) is used.

Coronary heart disease mortality calculates (B1)* as follows:

$$\begin{aligned} \mu = & 11.2889 - 0.588 \times \log(\text{systolic blood pressure}) - 0.1367 \times \text{smoking} - 0.3448 \times \log(\text{total/} \\ & \text{high density lipoprotein cholesterol}) - \\ & 0.1237 \times \text{electrocardiographic left ventricular hypertrophy} - 0.944 \times \log(\text{age}) - \\ & 0.0474 \times \text{diabetes} \\ \sigma = \exp(2.9851 - 0.9142\mu) & \quad (B1) \end{aligned}$$

Coronary heart disease events calculate (B2) *

$$\begin{aligned}\mu = & 15.5303 - 0.9119x \log(\text{systolic blood pressure}) - 0.2767x \text{smoking} - 0.7181x \log(\text{total/} \\ & \text{high density lipoprotein cholesterol}) - \\ & 0.5865x \text{electrocardiographic left ventricular hypertrophy} - 1.4792x \log(\text{age}) - \\ & 0.1759x \text{diabetes} \\ \sigma = & \exp(-0.3155 - 0.2784 \times (\mu - 4.4181)) \quad (B2)\end{aligned}$$

Step 2: For both Coronary Heart Disease events and Coronary Heart Disease (CHD) death equations the calculation of predictive time factor (**u**) is:

$$u = (\log(10) - \mu) / \sigma \quad \text{In this expression the length of time for follow up is 10 years}$$

Step 3: The predicted probability is calculated as

$$p = 1 - \exp(-\exp(u))$$

Step 4: the Percentage Predictive Risk (PPR) is

$$PPR = 1 - \exp(-\exp(u)) \times 100$$

b. Framingham risk equations for CHD events (B2) of women within 10 years

For the purpose of comparison, the female components of the original Framingham equations from Anderson et al., (1991) was modified as below for the computation of the predictive percentage risk of female having CHD in 10 years.

Step 1: For coronary heart disease events are calculated as follows

$$\begin{aligned}\mu = & 15.5303 - 0.9119x \log(\text{systolic blood pressure}) - 0.2767x \text{smoking} - 0.7181x \log(\text{total/} \\ & \text{high density lipoprotein cholesterol}) - \\ & 0.5865x \text{electrocardiographic left ventricular hypertrophy} - [\log(\text{age}/74)]^2 - \\ & 0.3758x \text{diabetes} \quad (B1)\end{aligned}$$

Use the equation (B2) from Anderson et al., (1991) to calculate **m** for women and note that

a = μ . Therefore for women, the computation is:

For women, compute

$$m = a - 5.8549 + 1.8515x [\log(\text{age}/74)]^2 - 0.3758x \text{diabetes}$$

If the age and diabetes part of the equation are moved into equation (B2) above then the expression below is obtained for female as:-

$m = a - 5.8549 + 1.8515$ that can be expressed as $m = \mu - 5.8549 + 1.8515$.

Put the above expression into (B2) and note that $m = \sigma$ therefore the result is

$$\sigma = \exp(-0.3155 - 0.2784x(\mu - 4.4181)) \quad (B21)$$

Step 2: Then the result will be (B22) below

$$\sigma = \exp(-0.3155 - 0.2784x(\mu - 5.8549 + 1.8515)) \quad (B22)$$

Step 3: From the equation calculate the predictive time factor (u) as:

$$u = (\log(10) - \mu)/\sigma \text{ in this expression the length of time for follow up is 10 years}$$

Step 4 : The Predicted Probability Risk is then given as:

$$p = 1 - \exp(-\exp(u))$$

The Percentage Predictive Risk (PPR) is therefore

$$PPR = 1 - \exp(-\exp(u)) \times 100$$

The above formulae were coded in MATLAB and the percentage predictive risk calculated for each of the participants selected from the HSE, (2006) and the results are in Table 8.6.

8.2.3 The International version (Zgibor et al., 2006):

An International Task Force working on Coronary Heart Disease and prediction models for Type-One diabetes patients proposed the International Framingham equation (Zgibor et al., 2006). In this research the results from the CMAUT Prognosis Framework will be benchmarked against the International Framingham equation. The procedure used for the computation of the PPR is based on the International Task Force approach, which is illustrated below.

Step 1: Computation of the location (μ) parameter for Men and Women that is denoted X1;

For Men

$$X1 = 11.1122 - 1.4792x \log(\text{age}) - 0.9119x \log(\text{SBP}) - 0.7181x \log(\text{Total Chol}/\text{HDL Chol}) - 0.2767x \text{smoking} - 0.1759x \text{diabetes} - 0.5865x \text{LVH ECG};$$

For Women

$$X1 = 5.2573 + 1.8515x [\log(\text{age})/74]^2 - 0.9119x \log(\text{SBP}) - 0.7181x \log(\text{Total Chol}/\text{HDL Chol}) - 0.2767x \text{smoking} - 0.3785x \text{diabetes} - 0.5865x \text{LVH ECG};$$

The rest of the calculations for the determination of the percentage predictive risk are the same for both men and women:

Step 2:- computation of the predictive time factor (u), which is denoted as X2 is as follows:

For both men and women:

$$X2 = [-2.1155149 - X1] / \exp(-0.3155 - 0.2784 \times X1);$$

Step 3:- the probability of CHD risk taken place within 10 years or Percentage Predictive Risk (PPR) in 10 years' time is given as:-

$$PPR = 100 \times (1 - \exp[-\exp(X2)])$$

The above formulae were coded in MATLAB and the predictive percentage risk calculated for each of the participants selected from the HSE, (2006). The results of the computation for each of the 3645 participants are presented in Tables and graphical format in section 8.3.

- Observation on the PPR values computed from the three Framingham methods.

After the computation of the predictive percentage risk for the three Framingham methods, it was observed that all the three different approaches give approximately the same PPR results for each of the selected participants. The Table 8.6A below shows the PPR values for each of the selected participants. In Table 8.6A the results earmarked as USA_PR is based on the computation from the USA original Framingham equation model I. This is followed by the results from the equations created by the International Task Force INT_PR (aka model II), while the model III is based on the UK calculation, which has two components namely the UKMEN_PR for men only and UK_PR for both man and woman.

8.3 Simulation Results for the Framingham Equations versions I, II and III:

- The Simulation results, tables and figures for the Framingham equations are below;

The Tables and Figures in the section are the results of inputting the demographic and clinical data of each of the selected 3654 participants into the Framingham version equations models I, II and III.

The Table 8.6A contains the results of the calculated 10 years PPR values for each of the first 10 participants using the three versions of the Framingham equations.

Similarly, Table 8.6B, at the end of this Thesis contains the PPR results of the first 30 participants and the PPR results of the entire 3645 participants are in the Appendix 8 Table 8.6C in electronic format.

The Table 8.7A shows the results of the computation of TPR, FPR, LRP and LRN for the Framingham equation models I – II – III based on the PPR for 10 years for the first 10 participants of the 3645 data set. The Table 8.7B, at the end of this Thesis contains the first 30 participants' results and the results of the entire group are in the Appendix 8 Table 8.7C in electronic format.

Table 8.6A predicative percentage risks for 10 years for the first 30 participants based on Framingham equation model I – II – III (I – USA, II – International, III – UK)

Pserial no.	Grp	BpI	Age	Sex	Ethnic	USA_PR	INT_PR	UKMEN_P R	UK_PR
13,956,102.00	No	No	60	Women	White	17.58	17.58	0.00	17.58
63,535,102.00	Yes	Yes	30	Women	White	0.13	0.13	0.00	0.13
71,831,101.00	No	No	66	Women	White	22.55	22.55	0.00	22.55
34,031,101.00	No	No	84	Women	White	12.87	12.87	0.00	12.87
72,604,102.00	No	No	59	Women	White	2.82	2.82	0.00	2.82
13,008,101.00	Yes	Yes	50	Women	White	15.12	15.12	0.00	15.12
39,139,101.00	No	No	34	Women	White	0.40	0.40	0.00	0.40
47,856,102.00	No	No	51	Women	White	1.22	1.22	0.00	1.22
37,710,101.00	No	No	61	Women	White	7.02	7.02	0.00	7.02
54,256,101.00	No	No	31	Women	White	0.05	0.05	0.00	0.05

Table 8.7A Calculation of TPR, FPR, LRP, and LRN, for the for the first 30 participants based on Framingham equation model I – II – III (I – USA, II – International, III – UK)

Pserial no.	Grp	BpI	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRN
13,956,102.00	No	No	60	Women	17.58	0	1	1	1.00	2808.99	0
63,535,102.00	Yes	Yes	30	Women	0.13	0	1	1	1.00	1406.47	0
71,831,101.00	No	No	66	Women	22.55	1	0	0.9988	0.9993	1404.7820	0.0012
34,031,101.00	No	No	84	Women	12.87	0	1	0.9988	0.9989	936.0825	0.0012
72,604,102.00	No	No	59	Women	2.82	0	1	0.9988	0.9986	702.3910	0.0012
13,008,101.00	Yes	Yes	50	Women	15.12	0	1	0.9988	0.9982	561.7548	0.0012
39,139,101.00	No	No	34	Women	0.40	0	1	0.9988	0.9979	468.0412	0.0012
47,856,102.00	No	No	51	Women	1.22	0	1	0.9988	0.9975	401.2857	0.0012
37,710,101.00	No	No	61	Women	7.02	0	1	0.9988	0.9972	351.0721	0.0012
54,256,101.00	No	No	31	Women	0.05	0	1	0.9988	0.9968	312.0275	0.0012

Figure 8.12 shows the graph of the computed 10 years PPR values of each of the 3645 participants against their individual PIND. The data used for plotting the graph is from Table 8.6C. Figure 8.13 is the prediction accuracy graph of the computed results of TPR against the FPR values of each of the 3645 participants.

This data used to plot the AUC graph is from Table 8.7C. The AUC areas of Framingham equation models I-USA, II-International and III -UK were calculated, by summing all the PPR data points using the trapezoidal method. The actual area is obtained by subtracting the sum of all the PPR data points from the sum of all the diagonal reference data points as discussed in chapter 3.

Figure 8.14 shows the graph of the discriminatory accuracy or discriminatory ability of the Framingham equation models. This was constructed by plotting the sensitivity and selectivity values of each of the 3645 participants in Table 8.7C. The graph of the sensitivity and selectivity against the NICE, (2006) recommended criterion and the interception are discussed in chapter 9. Table 8.7C contains the results of the computation of the positive and the negative Likelihood ratios for all the 3645 participants. Figure 8.15 is the performance accuracy graph of the PPR values of each participant obtained from the Framingham equations. This is done by plotting the positive and negative Likelihood ratio values on the Y-axis and the PIND of each participant on the X-axis. The graph is discussed in chapter 9.

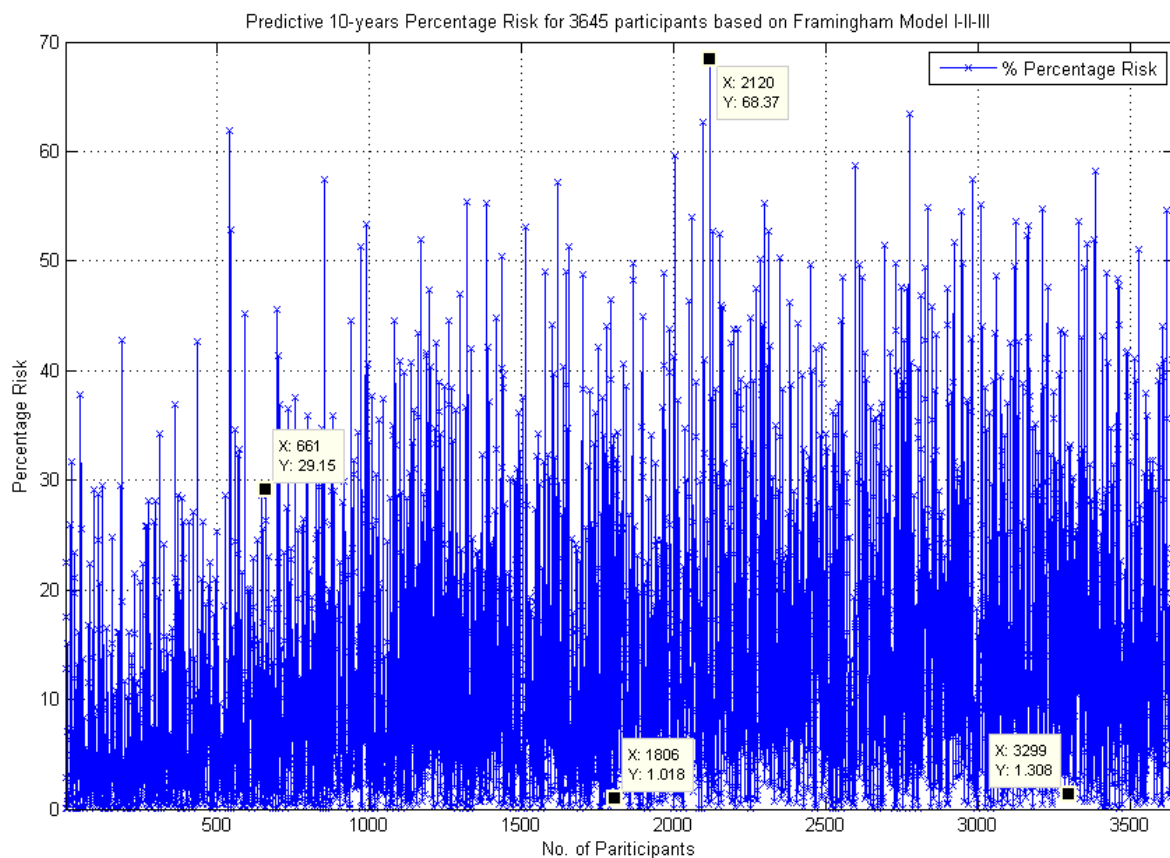


Figure 8.12: Predictive Percentage Risks for 10-years for Framingham I – II – III

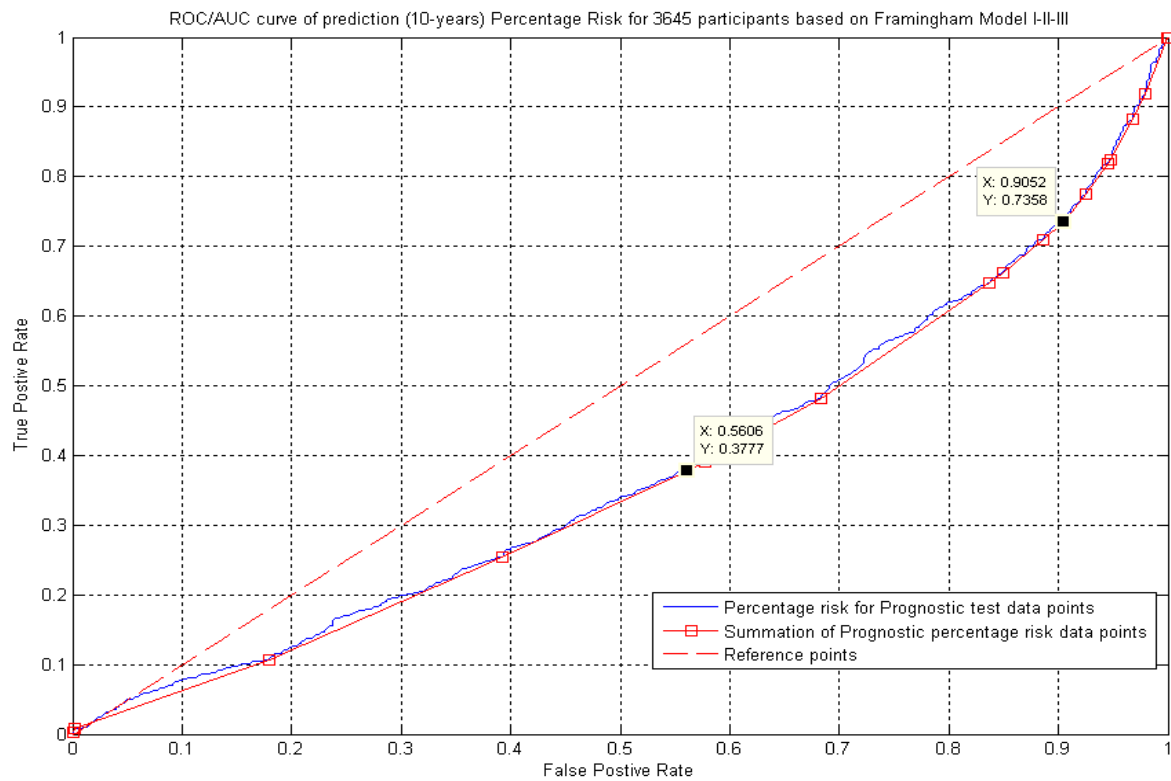


Figure 8.13: ROC/AUC for Framingham I – II – III

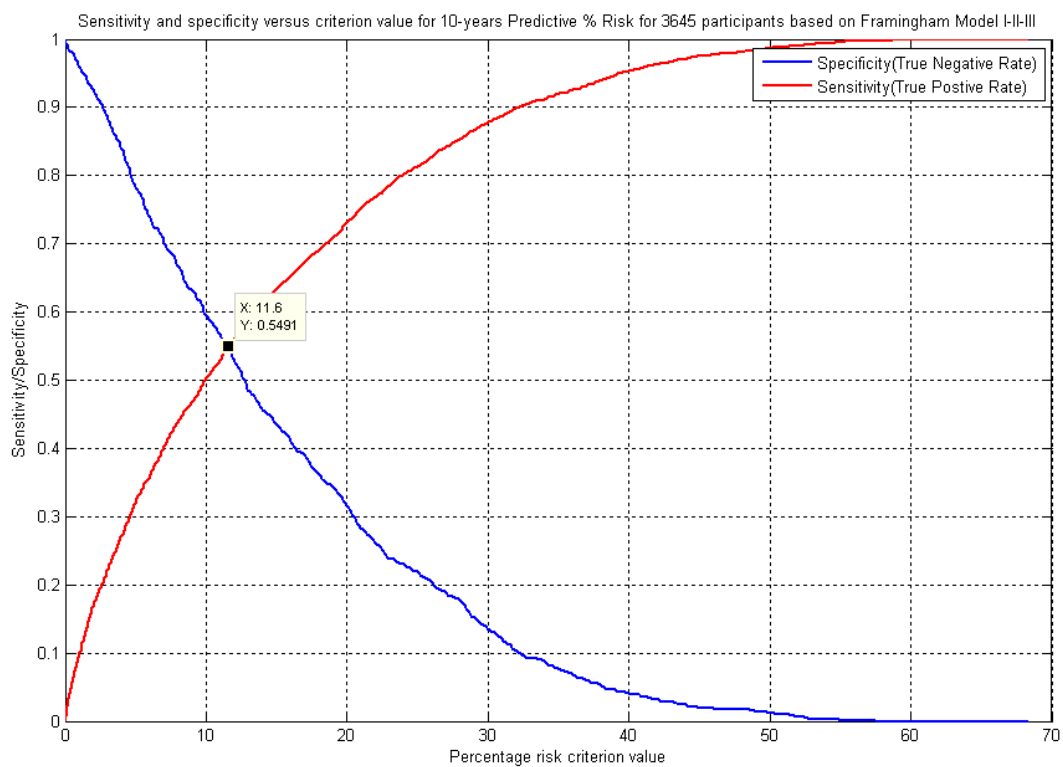


Figure 8.14: Sensitivity and specificity for Framingham I – II – III

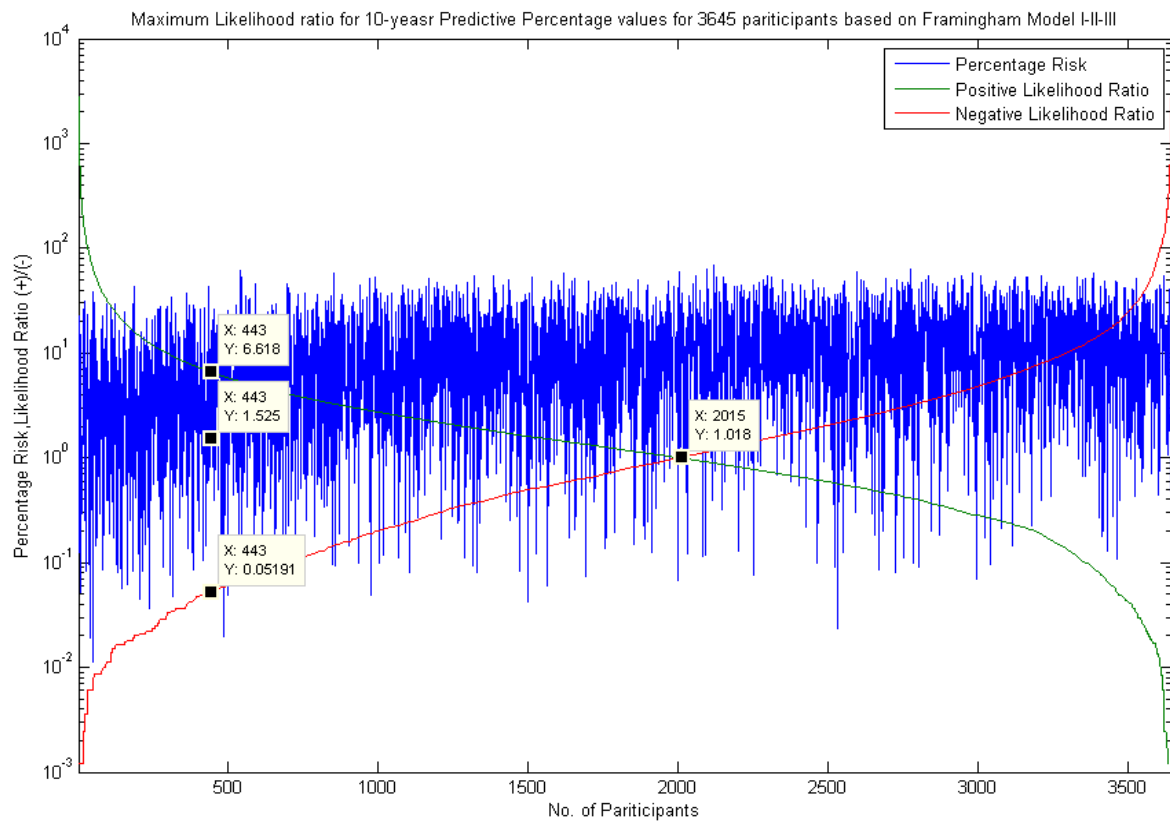


Figure 8.15: Maximum Likelihood ratio for Framingham I – II – III

8.4 Summary

In this chapter 8, two different approaches were used to determine the predictive percentage Risk (PPR) values of the participants been hypertensive. The first approach was the use of Web based CVD risk calculators, which were designed using the Framingham algorithm to compute PPR. The two Web calculators, NHS BlackHeath centre and Patient.co were used. The data sets of the 3645 selected participants were entered into each of the CVD risk calculators and the PPR values calculated for each participant.

The second approach is the use of Framingham equation to calculate the PPR values of the participants been hypertensive. The three different approaches used for the implementation of the Framingham Algorithm are the original Framingham Algorithm proposed by (Anderson et al, 1991), the England version by (Brindle et al., 2003) and the International Task Force version (Zgibor et al., 2006).

The data of selected 3645 participants were input into the MATLAB program, which were designed for each of the three Framingham equations for implementation purposes. It was observed from the results of the computed PPR values, that the three approaches have approximately the same PPR value for each of the 3645 participants. For the epidemiological analysis, the PPR values of each participant from the CMAUT Prognosis Framework were benchmarked against the web-based CVD risk calculators and the Framingham equations. The results were presented in tabular and graphical formats and they are discussed in the evaluation Chapter 9.

Chapter 9: Evaluation and Discussion

9.0 Introduction

This chapter starts with an evaluation of the Literature Review. This is followed by a summary of the different success criteria used to prove or disprove the hypothesis of this research as described in chapter 3. To achieve this goal, the results from the simulations carried out for each of the 3645 participants using the frameworks in chapters 5, 6, 7 and 8 were analysed and discussed. Then each of the success criteria was benchmarked against the results in chapter 8 and conclusion drawn. Finally, the two main component parts of the hypothesis were presented, discussed and inferences made.

9.1 Evaluation of Literature Review

From the literature reviewed, the research gap identified in CIS and CDSS is the need to reduce information overload by optimising the CVD clinical data for decision making. This research gap exists because different optimisation techniques have been used to address information overload in CIS but they all have limitations (Bertot, 2013).

The clinical data re-representation techniques introduced to address information overload in CIS are FOL –ERD and EAV/CR (Abu-Hanna et al, 2004) (Nadkarni et al., 2001). However, according to Roland (2000), EAV/CR technique has low performance rate as compared to the traditional conventional database and for its operation the SQL queries are complex and therefore it is difficult to implement and use (Nadkarni et al., 2006). Again, these two techniques do not have optimisation mechanisms in them therefore they cannot address the problem of information overload in CIS.

Therefore this research proposes an alternative solution, which is the application of data re-representation using UML class model that is formalised with the CMAUT and logical connectors. Based on the proposed solution the following hypothesis was established “that clinical data can be modelled using class diagram and re-represented with CMAUT logical connectors to reduce the space complexity in CIS and seamless converted them into the mathematical format that is optimised using LP algorithm to determine the risk of users been hypertensive”. To prove the hypothesis the following success criteria were used.

The criteria used are the Absolute Percentage Risk (APR) and Predictive Percentage Risk (PPR); computation of Prevalence and Kappa values; determination of space complexity, data size analysis using P-value before and after optimisation; calculation of sensitivity and selectivity; determination of the Accuracy of prediction models using AUC/ROC and Likelihood Ratio (Weinstein et al., 2001) (Fenton, 1997):

9.2 Clinical data representation using UML and CMAUT - Success criteria 1

To confirm that there is a relationship between clinical data and multiple attributes, the organs in the diseases domain were modelled with UML class diagram. In chapter 4, it was established that clinical data can be captured using class model and expressed in CMAUT format, which is seamlessly written in mathematical function. The generic expression for diseases, which affect complementary organs are written using the AND logic while diseases that are associated with substitutable organs use the OR logical connectors.

The disease associated with substitutable organs that uses OR expressions is kidney. The kidney was used as an example of substitutable organs because two kidneys can replace each other and have many attributes. Therefore the kidney system is represented in UML class model as a superclass with two subclasses and expressed in CMAUT using the OR connectors as $X_2 \equiv [(K_1 \vee K_2), P_1, S_1, P_2, S_2]$. In this expression K_1 and K_2 are the organs and P_1, S_1, P_2, S_2 are the attributes. The generic expression for these substitutable organs with multiple attributes is written as follows $[(C_1 \vee C_2, \dots, \vee C_n), P_1, S_1, \dots, P_n, S_n]$.

The disease associated with complementary organs and uses the AND expressions is CVD. Complementary organs are modelled in class diagram as unidirectional relation, which indicates that each class complements the other. The relationship between the classes is linked to the other class using the association arrow. In CMAUT data re-representation, the complementary organs with multi-attributes are expressed using logic connector AND as: $X_1: [(H_1 \wedge K_2 \wedge B_3), P_1, S_1, P_2, S_2, P_3, S_3]$: In this expression the CVD disease X_1 affects the body parts H_1, K_2 and B_3 where the organ H_1 has the attributes P_1 , and S_1 , whereas organ K_2 has attributes P_2 and S_2 etc. The generic expression for combinatorial clinical organs with multiple attributes using AND is $[(C_1 \wedge C_2, \dots, \wedge C_n), P_1, S_1, \dots, P_n, S_n]$. The logical expressions serve as the input to the optimisation framework that was discussed in chapter 5.

Convention of multiple attribute values into Utility Units: A disease X_1 that affects three organs can be expressed in CMAUT as $X_1: [(G_1 \wedge G_2 \wedge G_3), P_1, D_1, P_2, D_2, P_3, D_3]$. To convert the attributes in the expression into utility function (U) the formula used is $U = \sum[w_i f(s)]f(s) = \frac{P_i - P_o}{P_o}$. In this formula, the P_o is the expected blood pressure and the participant's measured blood pressure is P_i . The utility unit U_i of the organ G_i in the expression calculated as: $U_i = \sum(wpi f(spi) + wsi f(sdi) + wqi f(sqi))$. Therefore the disease X_1 is written in CMAUT format as $X_1: [(G_1 \wedge G_2 \wedge G_3), U_1, U_2, U_3]$.

The hypothesis of this research states that the application of CMAUT Data Re-representation will reduce the space complexity in the CMAUT Optimisation CVD framework as compared to the traditional non-CMAUT data representation technique used in FOL-ERD and EAV. This was discussed in chapter 4.

9.2.1 CMAUT Diagnosis Framework for CVD Risk Prediction -Success criteria 2

In chapter 5, two models of the CMAUT framework were designed and implemented to determine the initial clinical absolute percentage risk (APR) of a user having hypertension disease for diagnosis purpose. In this research, APR is defined as the probability of a participant having CVD based on their current multiple attributes values that are measured during medical examination. CMAUT framework model 1 was developed using beta coefficients (aka weights) obtained from conducting binary regression in SPSS using 4165 participants. The model 2 was developed using clinical data of the 3645 participants who are over 30 years. The beta coefficient values for the measurable multiple attributes were used to construct objective function that serve as input for the CMAUT algorithm. The objective function was optimised using the MATLAB linear programming technique to determine the absolute percentage risk of been hypertensive. This research used the CVD attribute values of the 3645 selected participants who provided full data during the HSE, (2006) survey.

The CMAUT Diagnosis Framework was used to determine the APR value of all the selected 3645 participants been hypertension YES or NO. This was done by first determining the percentage risk value of each participant and then comparing the results with the NICE, (2006) recommended value of 20%. The criterion states that everyone who has APR value of 20% or more percentage predicted risk value is prone to hypertension in future.

The validation of CMAUT Framework models 1 and 2 was conducted by comparing the absolute percentage risk results of the new Framework Diagnosis models with the predicted hypertension YES or NO results from the GPs in the HSE, (2006) report.

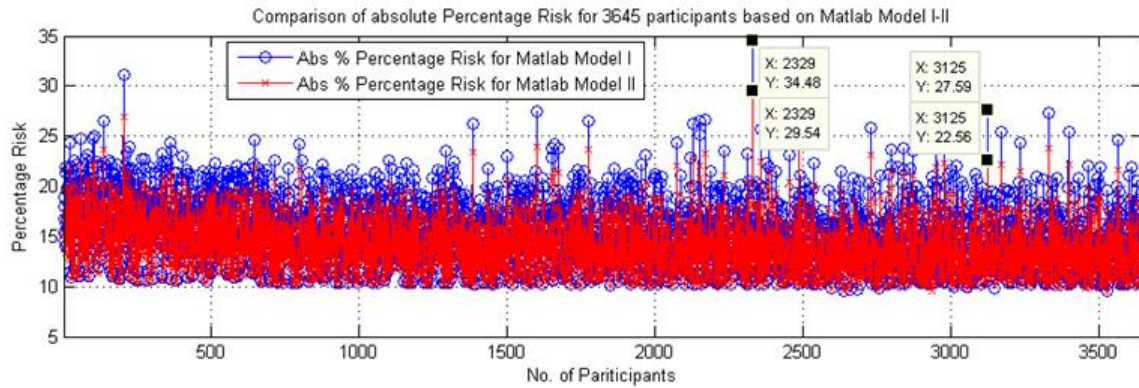


Figure 9.1: Comparison of Absolute Percentage Risk for 3645 participants based on CMAUT Diagnosis Framework Model I-II

9.2.2 Comparison of APR for CMAUT Model 1 and 2:

The relevance and kappa statistic of the two proposed CMAUT framework models were discussed in chapter 5. In this chapter 9, the comparative results of the CMAUT Diagnosis Framework model 1 and model 2 are recorded in Table 9.4C. For the purpose of comparison, the results of the simulation are for only the selected 3645 participants, who are over 30 years. The APR results from the CMAUT models 1 and 2 are benchmarked against the Prognosis Framework, Web Risk Calculators and the Framingham equations results in chapters 7 and 8. Figure 9.1 and Table 9.4C are summarised graph and tables that depict the comparison between the APR values for models 1 and 2. Table 9.4A is the comparative Table of the first 10 participants and the Table 9.4B in Appendix 9 contains the first 30 participants' results while the full results are on electronic format in Table 9.4C.

Table 9.4B: Comparison of Absolute Percentage Risk values from CMAUT models 1 and 2 using the first 10 participants:

Pserial no.	Grp	Bp1	Age	Sex	%PR(M-I Abso)	%PR(M-II Abso)	%PR(M-I Pre)	%PR(M-II Pre)	%PR(I-I Pr)	%PR(I-II Pre)	%PR(F-I P)
13,956,102.00	No	No	60	Women	14.6	13.8	15.6	14.79	7	9	17.58
63,535,102.00	Yes	Yes	30	Women	15.4	14.6	16.3	15.53	1	4.3	0.13
71,831,101.00	No	No	66	Women	16.7	15.4	17.7	16.42	8	20.9	22.55
34,031,101.00	No	No	84	Women	18.0	16.6	19.0	17.64	3	17.5	12.87
72,604,102.00	No	No	59	Women	18.9	17.3	19.8	18.26	3	11.8	2.82
31,510,102.00	No	No	20	Women	20.1	15.2	21.0	16.13	1	9.3	15.12
18,633,105.00	No	No	16	Women	18.5	19.5	19.4	20.28	2	8.7	0.40
13,008,101.00	Yes	Yes	50	Women	16.3	19.9	17.2	20.75	11	10	1.22
60,417,102.00	No	No	16	Men	17.4	14.3	18.3	15.26	0	7.5	7.02
39,139,101.00	No	No	34	Women	21.4	13.3	22.2	14.26	1	4.8	0.05

For comparison reasons, the clinical APR values of each of the 3645 participants for model 1 was superimposed on the graph of the APR results from model 2, which contains the same participants. The Figure 9.1 shows that the APR values of the participants in model 1 are higher than those in model 2. For example participant number 2329 in model 1 is 34.48% whereas in model 2 the value is 29.52%. Hence it is subsumed that the APR values of model 1 are higher than those from model 2.

- Kappa and Prevalence of CMAUT Diagnosis model 1 and 2 and GP Diagnosis

In chapter 5 section 5.6, it was established that the kappa value for CMAUT model 1 is $k=0.1508$ while the kappa value for CMAUT model 2 is 0.42. Using the Kappa Interpretation Table in Viera, (2005) and Cunningham et al., (2009), the $k = 0.1508$ means the APR values from CMAUT model 1 slightly agree with the actual GP diagnosed hypertension YES or NO. The CMAUT model 2 has k value of 0.42, which denotes that the APR risk values from the framework model 2 fairly agrees with the actual GP diagnosed hypertensive YES and NO.

In Chapter 5.6, the computed Prevalence for CMAUT model1 is 0.1857 while the Prevalence for CMAUT model2 is 0.0103. In percentages, the model 1 has a prevalence value of 18, 57% of the 3645 participants which denotes that the 677 participants were diagnosed as hypertension YES while the 2968 were hypertension NO. The model 2 Prevalence is 1.01 %, which means only 36 of the 3645 participants was identified as hypertension YES while 3609 participants were hypertension NO.

The prevalence from the HSE, (2006) report is 0.1857 or 18.57%, which means 677 participants were diagnosed as hypertension YES whereas 2968 were hypertension NO. Therefore, it is subsumed that the prevalence of model 1 is the same as the actual results from the HSE report as compared to model 2, which is lower than the actual prevalence. Based on these results the kappa statistic for model 2 is comparative better but the model must be enhanced to obtain a better prevalence.

9.3 Performance Evaluation of CMAUT Optimisation Framework

The third success criterion was used to verify the reduction of space complexity in the CMAUT optimisation framework. The technique used to measure the performance of space complexity of the optimisation framework is the big O function (Fenton et al, 1997). The space complexity analysis was conducted using mathematical operations that identify the relationship between the numbers of organs in the substitutable and complementary combinatory. The outcome was compared with the numbers of inequalities generated for their storage.

9.3.1 Space Complexity Analysis of CMAUT and Non-CMAUT CIS

As discussed in chapter 6, the Raman et al., (1999) mathematical technique was used to generate the set of inequalities for different numbers of substitutable and complementary organs in the disease domain. The mathematical operation was conducted for the CIS that uses the new CMAUT clinical data representation technique. The numbers of inequalities generated for each combinatory were recorded in Table 6.2. The same mathematical operation was repeated for CIS that use the traditional clinical data representation (aka Non-CMAUT CIS) such as EAV/CR and ERD-FOL discussed in chapter 4.

The numbers of constraints generated by the Non-CMAUT system for each combinatory were also recorded in Table 6.3. Table 9.1A is the summary of the constraints generated by both CMAUT and Non-CMAUT system for substitutable and complementary combination of organs. Figures 9.2 and 9.3 are the comparative graphs, which were plotted from the data in the Table 9.1C below. The full data is in appendix 9.1C.

Table 9.1A: No. of CMAUT and Non-CMAUT constraints for substitutable and complementary organs

No of organs (x) in the (OR)/(AND) combinatorial	No of constraints (inequalities) (y) for AND with CMAUT	No of constraints (inequalities) (y) for OR with and without CMAUT	No of constraints (inequalities) (y) for AND with Non- CMAUT
2	3	4	4
3	4	8	6
4	5	16	8
5	6	32	10
6	7	64	12
7	8	128	14
8	9	256	16

- Interpretation of space complexity results for CMAUT and Non-CMAUT CIS

From Table 9.1A and Figure 9.2, it is subsumed that when the numbers of complementary organs in the disease domain are increased, the numbers of inequality expressions generated for storage by both CMAUT and Non-CMAUT CIS also increases. Therefore, for CISs that use the CMAUT data re-representation the numbers of inequality expressions generated increases as the function $y = x + 1$ as shown in blue colour in Figure 9.2. The increase can be written in space complexity performance function using the big O notation as $O(x + 1)$. Again, from Table 9.1A and Figure 9.2, it is evident that for non-CMAUT CIS as the numbers of complementary organs in the combinatory increases the numbers of inequality expressions generated increases as a linear function $y = 2x$ as shown in red colour in Figure 9.2. Similarly for non-CMAUT the increase can be expressed in big O notation as $O(2x)$.

Therefore it is deduced that the CISs that use CMAUT data re-representation technique can capture and represent clinical data with half the number of expressions required by the same amount of data for the traditional non-CMAUT system. The deduction is supported by the big O notation of $O(2x)$ for non-CMAUT and $O(x + 1)$ for the CMAUT. This deduction is evident in Figure 9.2, where the curve for non-CMAUT system indicates that they require twice the number of constraints as compared to the CMAUT data representation technique.

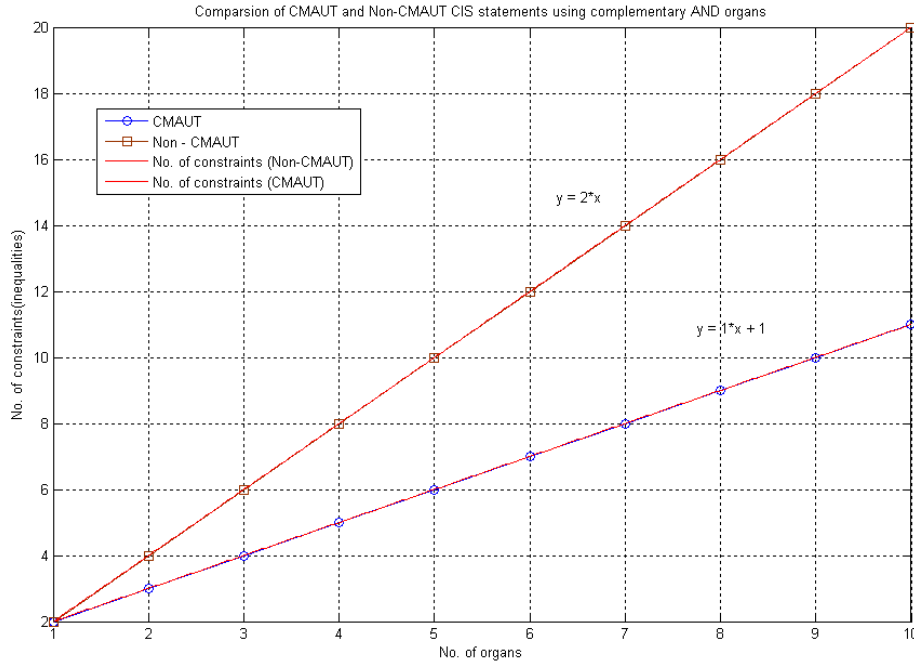


Figure: 9.2: Shows a comparison of *Non* – CMAUT and CMAUT CIS using complementary organs (*AND*).

For substitutable organs, the results from the mathematical operation for the different numbers of organs in the combinatory shown in Table 9.1B were used to plot the curves in Figure 9.3. From the inequality results in Table 9.1B and Figure 9.3, it is deduced that when the numbers of substitutable organs in the combinatory are increased the number of constraints also increases in the exponential function.

For both CMAUT and Non-CMAUT CIS the numbers of inequality expressions generated for substitutable organs increases with function 2^x , which is expressed in big O notation as $O(2^x)$. Figure 9.3 shows that both CIS data representations create the same numbers of inequality expressions as the numbers of substitutable organs are increased. In Figure 9.3 the CMAUT curve is shown in blue and the non-CMAUT data representation is in red, this means the number of constraints generated for both CMAUT and non CMAUT data representations are the same. Therefore for substitutable organs the storage space complexity for CMAUT and Non-CMAUT CIS are the same.

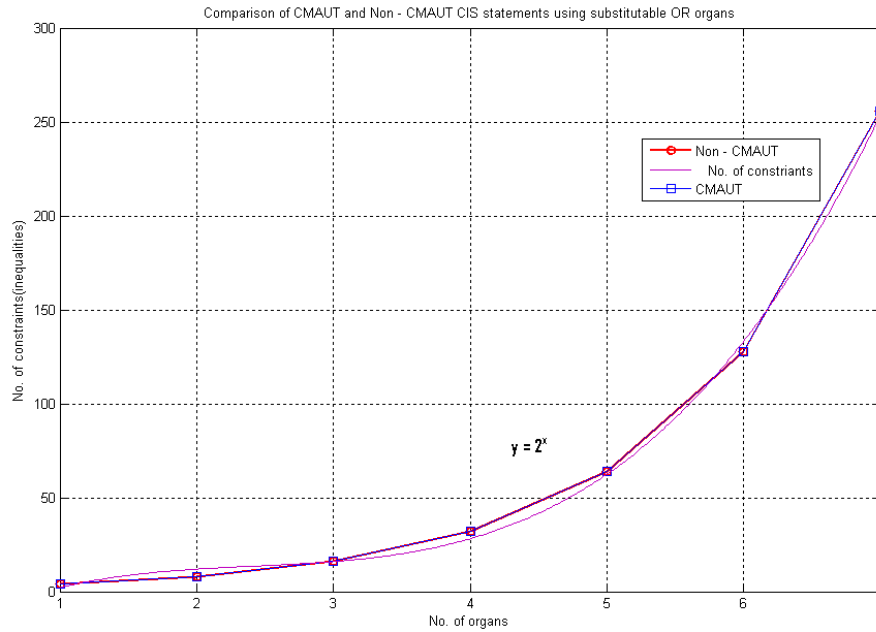


Figure 9.3: Shows comparison of Non- CMAUT and CMAUT CIS using substitutable organs (OR)

9.3.2 Clinical Data Sizes and T-test analysis of CMAUT Optimisation Framework

The second criterion is the comparison of the data sizes when clinical data are optimised with the CMAUT optimisation diagnosis framework as against when the data are not optimised. This criterion was used to prove the hypothesis, which states that the application of CMAUT framework will reduce the amount of clinical data required for decision making as well as the reduce space complexity in CIS. A cohort size of 402 participants were used for this analysis, which is approximately 10% of the 3456 participants that gave complete CVD clinical data and are over 30 years of age. The procedure used and the results obtained were discussed in chapter 6 and shown in Table 9.2A below.

First the CMAUT framework was used to determine the CVD Absolute Percentage Risk (APR) and the comparative attribute values of each of the 402 participants. In this context, the comparative attributes values are the output variables for each attribute and it indicates the results of comparing the recorded participant's attribute values with the standard expected values. The data sizes for each of the 402 participants were captured and measured before the optimization process and after optimizing the data with the CMAUT diagnosis framework. The results are in chapter 6 and Table 9.2A is the comparative Table with the results of the first 10 participants and Table 9.2B in the Appendix contains the results of first 30 participants while the full results are on electronic format in Table 9.2C.

Table 9.2A: Data size for first 10 participants before and after optimisation with CMAUT Framework

No. Of participants	PSerial no.	Data size before optimisation (bytes)	Data size after optimisation (bytes)
1	10,902,101.00	1256	465
2	10,846,103.00	1251	464
3	11,039,102.00	1251	463
4	11,046,101.00	1251	465
5	11,239,101.00	1245	464
6	11,249,102.00	1244	464
7	11,306,101.00	1249	464
8	11,313,101.00	1245	463
9	11,349,102.00	1243	464
10	11,356,101.00	1262	464

- Interpretation of the data sizes results from CMAUT and Non-CMAUT CIS

The data sizes obtained for each of the 402 participants before optimisation and after optimisation with the CMAUT framework and recorded in Table 9.2C were used to plot the curves in Figure 9.4 below. In Figure 9.4, the measured and recorded data sizes for each of the 402 participants before optimisation are shown in red while the data sizes after optimisation are shown in blue. It is deduced from the Table 9.2C that the average data size for the participant before optimization is approximately 1250 Bytes and the average data size after optimization with the CMAUT framework is 450 Bytes. This denotes that the average difference in data sizes between the two scenarios is 700 Bytes.

Figure 9.4 shows the difference in the data sizes where the data sizes without the CMAUT is in red line and the data sizes after optimisation with CMAUT framework are shown in blue. The difference in data sizes before and after optimization is proved below using the T-test statistical method.

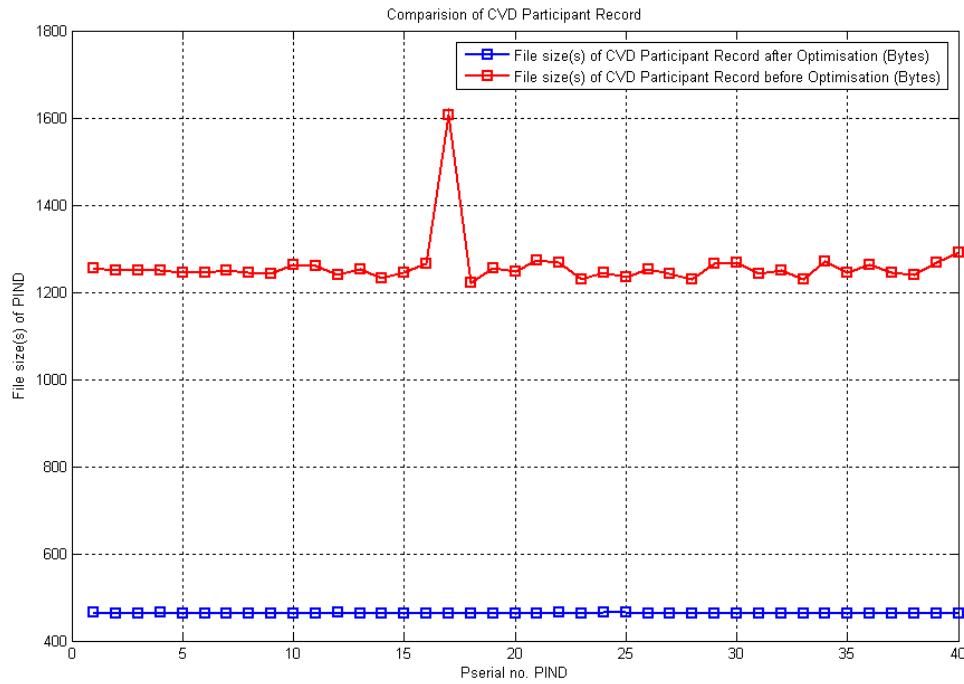


Figure 9.4: Before (red curve) and after optimisation (blue) of patient record data file

- Clinical Data Size and hypothesis statistical analysis using T-test

The pair T-test analysis was conducted, to prove the hypothesis that the use of the CMAUT Optimisation framework will reduce the clinical data sizes and space complexity required for decision making. The pair T-test was used because in the research two dependant sample data sizes of the same participants were compared at different period which are before and after optimisation. This is different from the independent T-test, which compares the samples from two different groups of participants (Campbell et al., 2006).

The CVD data of the selected 402 participants were used to perform the pair T-test statistical analyses in order to determine the confidence interval (CI) and p-value of the data sizes before and after optimisation. The analysis was conducted using the SPSS version 15 software. First, the data size of each of the 402 participants was retrieved from the HSE, (2006) and the data sizes before optimisation were measured and recorded Table 9.2. Secondly, the data sizes after optimisation using the CMAUT framework were measured and recorded. The Table 9.2C in Appendix 9 contains the entire data sizes results. Lastly, using the data in the Excel sheet Table 9.2C, the p-value and CI of the model were determined with the SPSS software. The result of the dependant pair T-test is shown in section 6.24.

Table 9.3 The Output of the Samples Statistics Results for the Paired Samples T-Test:

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	File size before Optimisation (In bytes)	1216.66	402	21.801	1.087
	File size after Optimisation (In bytes)	463.50	402	0.916	.046

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Size before Optimisation size after Optimisation (In bytes)	753.162	21.847	1.090	751.020	755.304	691.219	401	.000

From Table 9.3, it is observed that the mean values of the data sizes before optimisation and after optimisation are 1216.66 bytes and 463.50 bytes. The standard deviation of the data sizes before optimisation is 21.801 and after optimisation with CMAUT Framework is 0.916. The mean data size of 1216.66 Bytes before optimisation indicates that for non-CMAUT CIS the entire clinical data for each participant must be retrieved for decision making. In Table 9.3, the standard deviation of the data size after optimisation using the CMAUT framework is 0.916. This means the data size required for investigation is less with minimum variation because specific data is retrieved for making decision.

The T-test analysis gave a P-value of 0.000, which is less than 0.05, which means that the results are statistically significant. Therefore the alternative hypothesis which states that clinical data can be optimised using CMAUT framework to reduce the amount of data required for primary care investigation and information overload is accepted. Figure 9.4 and the calculated difference between the maximum data sizes before and after optimisation of 753.16 Bytes confirm this hypothesis.

Therefore the new CMAUT optimization framework reduces the data size retrieved and transmitted for decision making by approximately 700 Bytes for each participant. Thus CMAUT reduces information overload and the information needed for decision making. This proves the hypothesis, which states that the use of the CMAUT Optimisation framework will reduce the clinical data sizes and space complexity required for decision making in CIS.

9.4 Risk Prediction with CMAUT Prognosis Framework - Success criteria 2

In chapter 7, the CMAUT framework, was remodelled to determine the predictive percentage risk of a user been hypertensive within a period of 10 years. Again, in this research, Predictive Percentage Risk was defined as the probability that a user will develop hypertension in a specified period of time based on their present measurable and non-measurable CVD risk factors. The CVD Risk factors used are Age, Sex, Heart Beat, BMI, systolic pressure, diastolic pressure, HDL Cholesterol, MAP blood pressure, Diabetic, Total Cholesterol, Smoking, Existing CVD and Existing ECG.

The first part of the optimisation algorithm in the Prognosis framework is used to determine the absolute percentage risk (APR) of a participant been hypertensive based on their measurable attributes. The second part is the computation of predictive time factor (u) in 10 years' time using both the measurable and non-measurable CVD risk factors. This was done by adapting the Weibull distribution and Framingham methods explained in Anderson et al., (1990). The Predictive Percentage Risk aka $P(T)$ is the arithmetical sum of the absolute percentage risk (APR) and the 10 years predictive time factor ($P(t)$). This Predictive Percentage Risk approach is similar to the algorithm used for all the three Framingham equations and also for the Web based CVD risk calculators.

9.4.1 Comparison of Predictive Percentage Risk for CMAUT Models 1 and 2:

The simulation results of the CMAUT Prognosis Framework model 1, which was designed with the data for the over 16 years' old participants and model 2 of over 30years old were recorded in Table 9.4B. The PPR results of the selected 3645 participants from the CMAUT Prognosis models 1 and 2 are recorded in Table 9.4B before they were benchmarked against the Web Risk Calculators and the Framingham equations.

For comparison, the PPR results of each of the 3645 participants for CMAUT Prognosis model 1 was superimposed on the graph that depicts the results of the PPR values from model 2 for the same participants. Figure 9.5 below is the summarised graphs that depict the comparison between PPR results from the CMAUT Prognosis models 1 and 2.

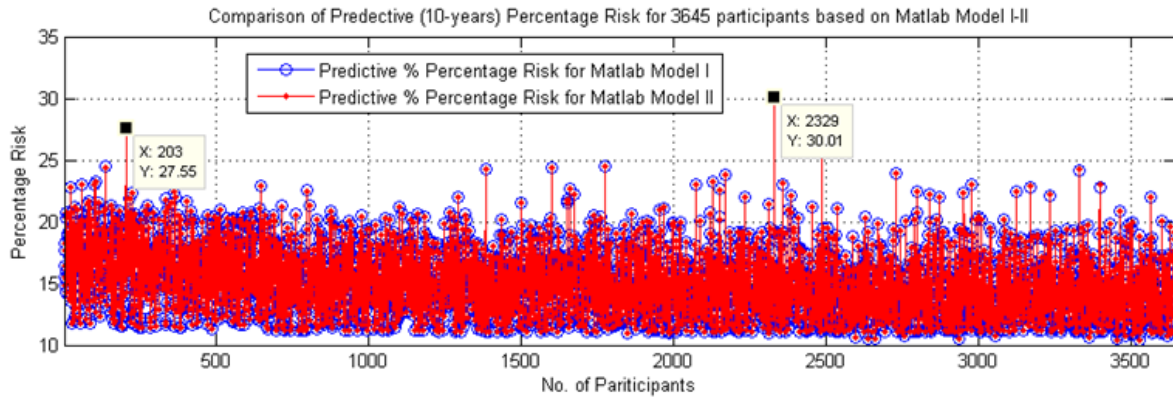


Figure 9.5: Comparison of Predictive Percentage Risk for 3645 participants based on CMAUT Prognosis Framework Model I-II

The Figure 9.5 shows that the PPR values of the selected participants in model 1 are higher than the PPR values in model 2. For example the PPR value for the participant number 2329 in model 1 is 35.28% while that from model 2 is 30.32%. Therefore it is subsumed that comparatively the PPR values for all participants in model 1 are lower than PPR values in model 2. For both Diagnosis and Prognosis CMAUT frameworks, the PPR values for model 1 are higher than model 2 for the same participant. It is subsumed that model 1 gives overestimated PPR values, which needs redress. The CMAUT PPR values are compared with the results from the Framingham equations and Web risk CVD calculators and they are discussed in the conclusion Chapter 10.

- Kappa Statistics and Prevalence for CMAUT Prognosis model 1 and 2 with GP

In chapter 5, section 5.6 the kappa statistics of model 1 and 2 were calculated by comparing the results of the APR values from the CMAUT diagnosis Framework with the GP diagnosis BP_YES and BP_NO of the HSE, (2006) report. It was established that Model 2 gives realistic APR values as compared to Model 1. From Figure 9.3 and Table 9.8C, it is concluded that Model 1 gives overestimated PPR values. Therefore emphasis must be placed on the analysis of CMAUT framework Model 2, Framingham equations and Web CVD risk calculators as CVD prediction models.

Inferring from the above, the CMAUT Prognosis Predictive model 2 results and the GP diagnosis BP_YES and BP_NO from the HSE, (2006) were used to calculate the kappa value for the 3645 selected participants. Below are the results:

Percentage _YES = 8.7149

Percentage _NO= 120.8221

Table 9.4G: the actual_agree_YES and NO in tabular form

	Yes	No	Total
Yes	677	2968	3645
No	59	3586	3645
	736	6554	7290

Using the information in Table 9.4G, the kappa value k is computed using the formula $k = \frac{P_o - P_e}{1 - P_e}$ where $P_e = [(n_1/n) * (m_1/n) + ((n_0/n) * (m_0/n))]$ and $P_o = [(a + d)/N]$ as shown in Chapter 5.12.

The calculated kappa value for the CMAUT Prognosis Predictive model II is 0.1695. According to Viera et al, (2005) and Cunningham, (2009), $kappa = 0.1695$ denotes that the PPR results from both models have a fair agreement between them. Hence, it is subsumed that the kappa value calculated using the GP hypertension diagnoses model and the CMAUT model 2 is 0.169, that is approximately 0.2. According to Viera, (2005), the kappa value of 0.2 means there is a slight agreement between the two prediction models.

The Prevalence of the CMAUT Prognosis Predictive model 2 and the GP diagnoses hypertension (YES and NO) was calculated using the data in Table 9.4C:

The GP diagnoses Prevalence was computed using GP YES and GP_NO as follows:

$$a/(a + b) = 677/(677 + 2968) = 677/3645 = 0.1857 \text{ or } 0.186 = 18.6\%$$

The CMAUT model 2 Prevalence was computed using the results from Table 9.4C, where the YES is 59 and NO is 3586: The model 2 Prevalence value was computed as follows:

$$a/(a + b) = 59/(59 + 3586) = 59/3645 = 0.01618 \text{ or } 0.0162 = 1.62\%$$

The prevalence value is the proportion of identified hypertension suffers among the selected 3645 participants as diagnosed by the GPs as 18.6%, which is higher than the 1.62% identified by the CMAUT model 2. From the results of the kappa statistic, it is subsumed that CMAUT model 2 should be enhanced to achieve better prevalence.

9.4.2. Benchmarking CMAUT Prognosis framework with other CVD Prediction Tools

To verify the results from the CMAUT Prognosis model 2, first the PPR results are compared with the results from the Blackheath and Patient UK websites, which are the two selected Internet based CVD calculators. In Chapter 8, the PPR results of each of the 3645 participants were determined using the two web based CVD calculators. The PPR values of each participant was converted into hypertension YES or hypertension NO using the NICE recommended PPR value of 20%. According to NICE, (2006), when the results of the calculated PPR value is greater or equal to 20%, the participant is considered as hypertension YES: Alternately, if the value is less than the 20 percentage then the participant is earmarked as hypertension NO. The PPR results from the CMAUT Prognoses models indicate the models 1 and 2 can be used as prediction models and as an epidemiological tool to determine the percentage risk of a user been hypertensive or not.

- Benchmarking CMAUT Prognosis Framework with CVD web risk calculators

In chapter 7, the CMAUT Prognosis Framework had incorporated in it an algorithm that calculates the PPR value of a participant been hypertensive in 10 years based on their current situation. This is similar to the CVD Web calculators, which are designed to predict risk of hypertension in 10 years' time. Therefore the PPR values from CMAUT framework in Chapter 7 were compared with the Internet based CVD calculators' results in Chapter 8.

The two Internet CVD Risk calculators that were selected and used are Internet model I: - NHS BlackHeath, <http://www.bhgp.co.uk/chdriskresult.asp> and Internet model II Patient UK (ref: <http://www.patient.co.uk/doctor/Primary-Cardiovascular-Risk-Calculator.htm>). These websites have now been updated and they use Qrisk®2 instead of the Framingham equations.

The comparison between NHS Blackheath – Internet model 1 and the Patient UK- Internet model 2, revealed that the (10-years) PPR values from Patient UK are lower than the NHS Blackheath values. For example the participant with PSerial number of 15415101 or PIND 1956 in Figure 9.6 was predicted to have 63% risk by NHS Blackheath website but Patient UK gave a low value of 13.4%.

Further analysis of the simulation results in Table 9.8C confirm that the PPR values from Patient UK are lower than the NHS Blackheath values see Table 9.8C in Appendix 9. This proves that there are inconsistencies in the PPR values calculated by the Internet CVD risk calculators. This led to the NICE, (2010) recommendation that Framingham based prediction tools used in UK give higher estimated PPR risk values.

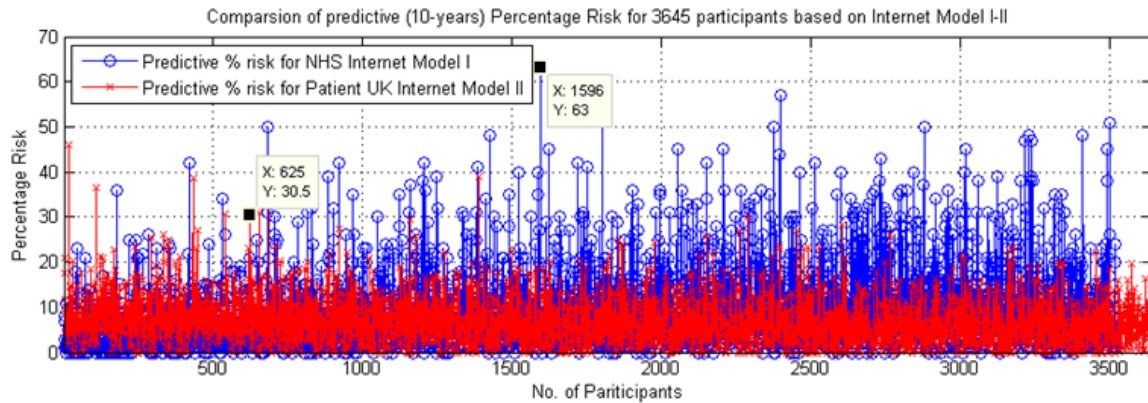


Figure 9.6: Comparison of (10-years) Predictive Percentage Risk for 3645 participants using Internet Calculators Model I-II

- Computation of kappa value for Internet model I and Internet Model II results:

To verify the PPR results from the two Internet CVD risk calculators, the kappa values for the two websites, Internet model I: - NHS BlackHeath, and Internet model II Patient UK were calculated and analysed as below:

Percentage _YES = 1.0050

Percentage _NO = 98.8807

Table 9.5: The actual_agree_YES and NO in tabular form

	Yes	No	Total
Yes	398	3127	3525
No	4	3092	3096
	402	6219	6621

The result of the computation using the formula $k = \frac{P_o - P_e}{1 - P_e}$ gives a kappa value of 0.1052.

The kappa value of 0.1052 indicates that there is a slight agreement between the PPR values

from the two Internet CVD risk calculators. The comparison of the risk values from the Internet CVD calculators and the Framingham equations are discussed in section 9.5, using the results shown in Appendix 8.

- Benchmarking CMAUT Prognosis Framework with Framingham Equations

The results from the proposed CMAUT Prognosis Framework were compared with the three selected CVD Framingham risk algorithms used in this research. The three Framingham equations are:

The original Framingham equation from the USA, which was used because it is still used as the basis for designing CVD risk prediction models even though according to NICE, (2010), the equation gives over estimate PPR values. The second equation used is from Brindle et al, (2003) because it was proposed for cohorts in England, UK. The third equation proposed by the International Task Force was developed for European countries (Zgibor et al., 2006).

The original Framingham CVD equations in Anderson et al., (1991) from USA, was used to calculate the PPR values for each of the 3645 participants and the results recorded in Table 9.8C. Secondly, since the cohorts and the participants used in this research are from England, UK, the (Brindle et al, 2003) was used for the male cohort. The Brindle et al., (2003) equation was modified in accordance with the original Framingham equation to cater for women in England (Brindle et al, 2006) and the modification was explained in Chapter 8. Finally the International equation proposed by Zgibor et al., (2006) was used to calculate the PPR values for each of the 3645 participants and the results recorded in Table 8.6C. Below is Table 9.8C and the graph that shows the comparative PPR results for the three Framingham equations the other results are shown in Appendix 9;

In chapter 8, it was observed that the calculated PPR values for all the 3645 participants using the three different Framingham equations were approximately the same values. For example in Figure 9.7, the PPR for the participant with PIND 2120 (PSerial 37213101) gave 66, 53% when the UK equation was used but give 68.36% when the USA/International was used. Similarly, the participant with PIND 2774 (PSerial of 79633101) had 63, 74% based on the UK equation but 63.37% when the USA and International equations were used.

In clinical terms, these PPR values are almost the same but according to NICE, (2006) CVD guidelines these people should be dead because their percentage risk values are three times higher than the recommended value of 20%. The differences in the PPR values from the three Framingham equations and the other CVD Prediction models are summarised below:

- For CMAUT model 2, the participant with PIND 2120 (PSerial 37213101) has PPR value of 16.439%, while the participant with PIND 2774 (PSerial of 79633101) has 10.263%.
- For Internet model 1, the participant with PIND 2120 (PSerial 37213101) has PPR value of 10%, while the participant with PIND 2774 (PSerial of 79633101) has 17%.
- For Internet model 2, the participant with PIND 2120 (PSerial 37213101) has PPR value of 8.2%, while the participant with PIND 2774 (PSerial of 79633101) has 5.1%.

From the above analysis, it is concluded that the three Framingham equations have the highest PPR values for the same participants, which are over 60%, whereas the CMAUT Prognosis framework model 2 has 16% and 10% respectively. The PPR values for the Internet model 1 are 10% and 17% while Internet model 2 has the lowest PPR values of 8% and 5% respectively. It is therefore subsumed that the PPR values for Internet model 1 and the CMAUT model 2 are comparable while the other CVD prediction models have different PPR values hence more research are required.

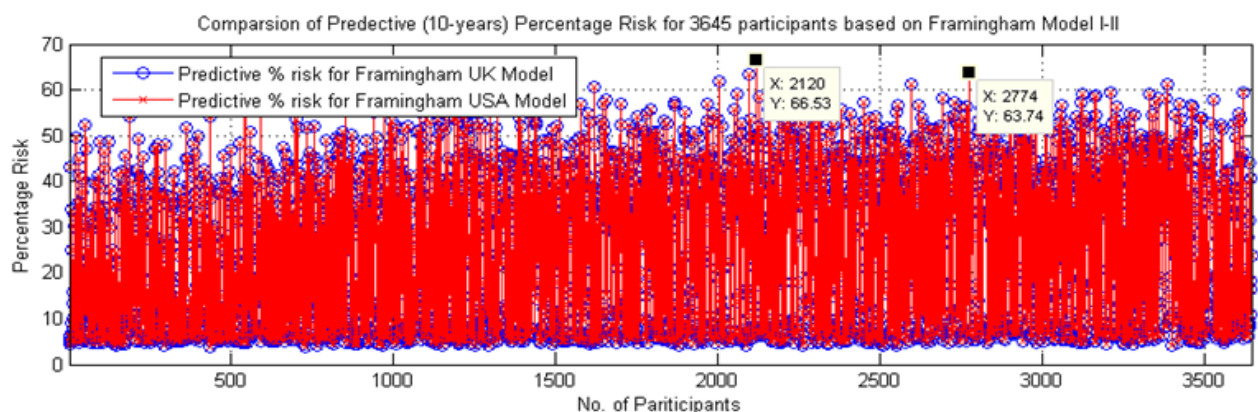


Figure 9.7: Comparison of (10-years) Predictive Percentage Risk for 3645 participants using Framingham Models 1-II

When a critical analysis of all the computed PPR values of the entire 3456 participants in Table 8.6C were conducted; it was revealed that the results from the original Framingham equation model I of USA and the International equation model II are the same but are different in some cases for the UK equation model III.

This is because all the three equations were developed and implemented using the same technique proposed by Anderson et al, (1991). The difference between the three equations is that the data of the CVD risk factor used for conducting the statistical binary logical regression analysis are from different countries. A review of the three equations shows that there are minor changes in the values of the constants in the equations but the method of computation is the same in all the three equations. It was established that the UK equation model III was original designed only for UK men but had to be modified in this research to include female. Therefore there are some differences between the PPR values from the USA model I and the UK equation model III. However, for analytical purposes, this research subsumes that the three Framingham equations are the same unless otherwise specified.

9.4.3 Computations of kappa value for Framingham equations I and II:

The kappa value is used to determine the level of agreement between two prediction models. In this section since the three Framingham equations have identical percentage risk values, the PPR values for the Framingham equation model I from USA and the International equation model II were grouped into hypertension YES and NO using the NICE, (2006) benchmark of 20%. The result of classifying the 3645 selected participants into YES and NO groups are presented below in Table 9.6. The computation of the kappa value for the USA equation model I and the International equation model II is as follows:

$$\text{Percentage_YES} = 68.1873$$

$$\text{Percentage_NO} = 49.5021$$

Table 9.6: The actual_agree_YES and NO in tabular form

	Yes	No	Total
Yes	833	2812	3645
No	568	1392	1960
	1401	4204	5605

The data in Table 9.6 and the formula $k = \frac{P_o - P_e}{1 - P_e}$ were used for the kappa computation and the result gives a kappa value of -0.0484.

This kappa value is approximately 0.05, which indicates that there is a good agreement between the USA equation model I and the International equation model II. However, the arithmetical sum of the Percentage _YES and Percentage _NO as well as negative kappa value are two issues that need further investigation and are not addressed in the research.

- Comparison of PPR values from Framingham equations with CMAUT model 2.

The PPR values from the three Framingham equations are identical and therefore the calculated kappa value of approximately 0.05 indicates a good level agreement between the different prediction models. Again, the PPR values from the Framingham equations were compared with the PPR values from the CMAUT framework model 2. The Framingham equation from USA and UK were grouped into hypertension YES and NO using the NICE, (2006) benchmark value of 20%. The outcome of classifying PPR values into the YES and NO groups were compared with the results from the CMAUT Prognosis Framework model 2. The Kappa value, which indicates the level of agreement between the CMAUT Prognosis model II and the Framingham Equation Model I UK_USA is shown:

Percentage _YES = 23.7288

Percentage _NO = 77.1612

Table 9.7: The for actual_agree_YES and NO in tabular form

	Yes	No	Total
Yes	59	3586	3645
No	14	2767	2781
	73	6353	6426

The information in Table 9.7 and the formula $k = \frac{P_o - P_e}{1 - P_e}$ were used for the computation of kappa and it gives a kappa value of 0.0097. The computed kappa value is approximately 0.01 indicates that there is a slight agreement between the CMAUT Prognosis model 2 and the Framingham Equation.

From the above analysis, it is subsumed that there are slight levels of agreement between the CMAUT Prognosis framework module 2, Framingham Equations and the NHS Blackheath

(Internet model 1). However, there are inconsistencies in the PPR values from the other prediction models, which are shown in Table 9.8A with more details in Appendix 9.

Table 9.8A: Comparison of 10-years PPR values from CMAUT models, Internet calculators and Framingham equations using the first 10 participants:

Pserial no.	Grp	Bp1	Age	Sex	%PR(M-I Absol)	%PR(M-II Abso)	%PR(M-I Pre)	%PR(M-II Pre)	%PR(I -I Pr)	%PR(I -II Pre)	%PR(F-I P)
13,956,102.00	No	No	60	Women	14.6	12.8	15.6	13.1	7	9	17.58
63,535,102.00	Yes	Yes	30	Women	15.4	14.4	16.3	15.4	1	4.3	0.13
71,831,101.00	No	No	66	Women	16.7	12.2	17.7	12.6	8	20.9	22.55
34,031,101.00	No	No	84	Women	18.0	14.4	19.0	15.3	3	17.5	12.87
72,604,102.00	No	No	59	Women	18.9	13.4	19.8	13.8	3	11.8	2.82
13,008,101.00	Yes	Yes	50	Women	20.1	13.5	21.0	14.5	1	9.3	15.12
39,139,101.00	No	No	34	Women	18.5	19.1	19.4	19.3	2	8.7	0.40
47,856,102.00	No	No	51	Women	16.3	19.4	17.2	19.6	11	10	1.22
37,710,101.00	No	No	61	Women	17.4	14.1	18.3	14.4	0	7.5	7.02
54,256,101.00	No	No	31	Women	21.4	11.7	22.2	12.1	1	4.8	0.05

From Table 9.8C, it is subsumed that they are difference between the various CVD risk calculators. This inconsistency led to the recommendation in NICE, (2010). Again some of the calculators give percentage risk values, which can be interpreted as the participant must die, while in reality they are alive. It is therefore recommend that a detail research must be conduct on the Web CVD calculator to address the issue of inconsistency in the PPR results. There must be standardisation of the input metrics used for developing risk calculator, adaptation of acceptable algorithm and all Web CVD calculators must be benchmarked.

9.5 Prediction models Accuracy with Sensitivity/Selectivity and AUC/ROC

In addition to benchmarking the CMAUT model 2 against existing Framingham equations and Web CVD calculators, the following analyses were also conducted. The accuracy of the different prediction models were determine using sensitivity/selectivity, AUC/ROC and Likelihood ratio in accordance with success criteria 6.

9.5.1 Discriminatory ability of three prediction models using sensitivity/selectivity

The CMAUT Prognosis model 1 and 2 were discussed and used to predict the PPR values of the 3645 selected participants in chapter 7. The discrimination accuracy of each of the two prediction models was determined by first calculating the sensitivity (aka TPR) and specificity or selectivity (aka FPR) values of each of the participant using the NICE, (2006)

recommended PPR criterion of 20 %. In this research, the terms specificity and selectivity are the same and are used interchangeably.

In Table 9.9A are the results of computing the TPR and FPR values for CMAUT Prognosis models, Internet calculators and Framingham equations for the first 10 participants. In Appendix 9 is Table 9.9B that contain the results of the first 30 participants and the results of the entire 3645 selected participants are in Table 9.9C in electronic format.

Table 9.9A: Comparison of TPR and FPR for CMAUT models, Internet calculators and Framingham equations of the first 10 participants:

Pserial no.	Grp	TPR(M-I)	FPR(M-I)	TPR(M-II)	FPR(M-II)	TPR(I-I)	FPR(I-I)	TPR(I-II)	FPR(I-II)	TPR(F-I)	FPR(F-I)
13,956,102.00	No	1	1.000	1	0.9997	1	0.99968	1	0.9997	1	1.00
63,535,102.00	Yes	1	0.999	1	0.9994	1	0.99936	1	0.9994	1	1.00
71,831,101.00	No	1	0.999	1	0.9991	1	0.999041	0.984848	0.9994	0.9988	0.9993
34,031,101.00	No	1	0.999	1	0.9989	1	0.998721	0.984848	0.9991	0.9988	0.9989
72,604,102.00	No	1	0.999	1	0.9986	1	0.998401	0.984848	0.9988	0.9988	0.9986
31,510,102.00	No	0.9983	0.9987	1	0.9983	1	0.998081	0.984848	0.9986	0.9988	0.9982
18,633,105.00	No	0.9983	0.9984	0.9935	0.9983	1	0.997761	0.984848	0.9983	0.9988	0.9979
13,008,101.00	Yes	0.9983	0.9981	0.9871	0.9983	1	0.997442	0.984848	0.9980	0.9988	0.9975
60,417,102.00	No	0.9983	0.9979	0.9871	0.9980	1	0.997122	0.984848	0.9977	0.9988	0.9972
39,139,101.00	No	0.9966	0.9979	0.9871	0.9977	1	0.996802	0.984848	0.9974	0.9988	0.9968

- Interpretation of sensitivity/selectivity decision plot

Using the results from Table 9.9C in Appendix 9, the sensitivity (TPR) and selectivity (TNR) for the CMAUT Prognosis model 1 and 2 were plotted against the selected criterion values. In this research, the percentage risk criterion value is the NICE, (2006) criterion of 20 %. The x: y interception values of the sensitivity and specificity curves show the degree of discrimination accuracy of the prediction models.

The discrimination decision graph assists in the selection of the optimum levels, which are the optimum PPR values for each of the model in terms of their sensitivity and selectivity. Again, in each graph the point of interception shows the optimum decision PPR value, where the maximum number of incidences of the CVD disease is correctly diagnosed or prognoses as positive (i.e. present) or negative (i.e. absent). This denotes that from the graphs, the interception point (x:y) is the decision criterion value that correctly identifies the proposition of the correctly identified TPR and TNR of the models under consideration.

From Figure 9.8 below, the sensitivity and specificity curves of model 1 intercept at the PPR percentage and sensitivity/specificity value of (14, 47%:0, 43). In model 2 the interception is at the PPR percentage and sensitivity/specificity value of (13,89% : 0, 58). This means that for model 1 the decision PPR criterion value is 14.47% and the proposition of TPR and TNR is 0, 43 or 43% while model 2 criterion value is 13, 89% with correct proposition of 0, 58 or 58%. Therefore it is subsumed that model 2 has approximately 58% of the prediction been correct. However, the PPR criterion value of 13, 89% is low as compared to model 1 of 14.47% and the NICE recommended percentage risk of 20%.

For the Web Based CVD prediction models, the sensitivity and specificity curves for Internet model 1 intercept at the PPR percentage and the sensitivity/specificity value of (8%:0,43) while the Internet model II intercepts at percentage and the sensitivity/specificity value of (5.7% : 0, 58). In contrast, the Framingham equations intercept at PPR percentage and the sensitivity/specificity value of (11.6%: 0, 549) as shown in Figure 9.8.

Therefore it is subsumed that the decision criterion values for the Web based CVD calculators which are 8%, and 5.7% are lower than the Framingham equation model PPR value of 11.6% and still lower than the NICE recommended value of 20%. However, the Web based CVD calculators correctly identify higher proposition of the TPR and TNR of 58, 9 % and 47% which are more than the three CVD Predictive models. Thus the Internet model II has a better discriminatory accuracy and proposition prediction ability as compared to all the other prediction models but its decision criterion value of 5.7% is nowhere near the CMAUT Prognosis models of 14% and the NICE value of 20%.

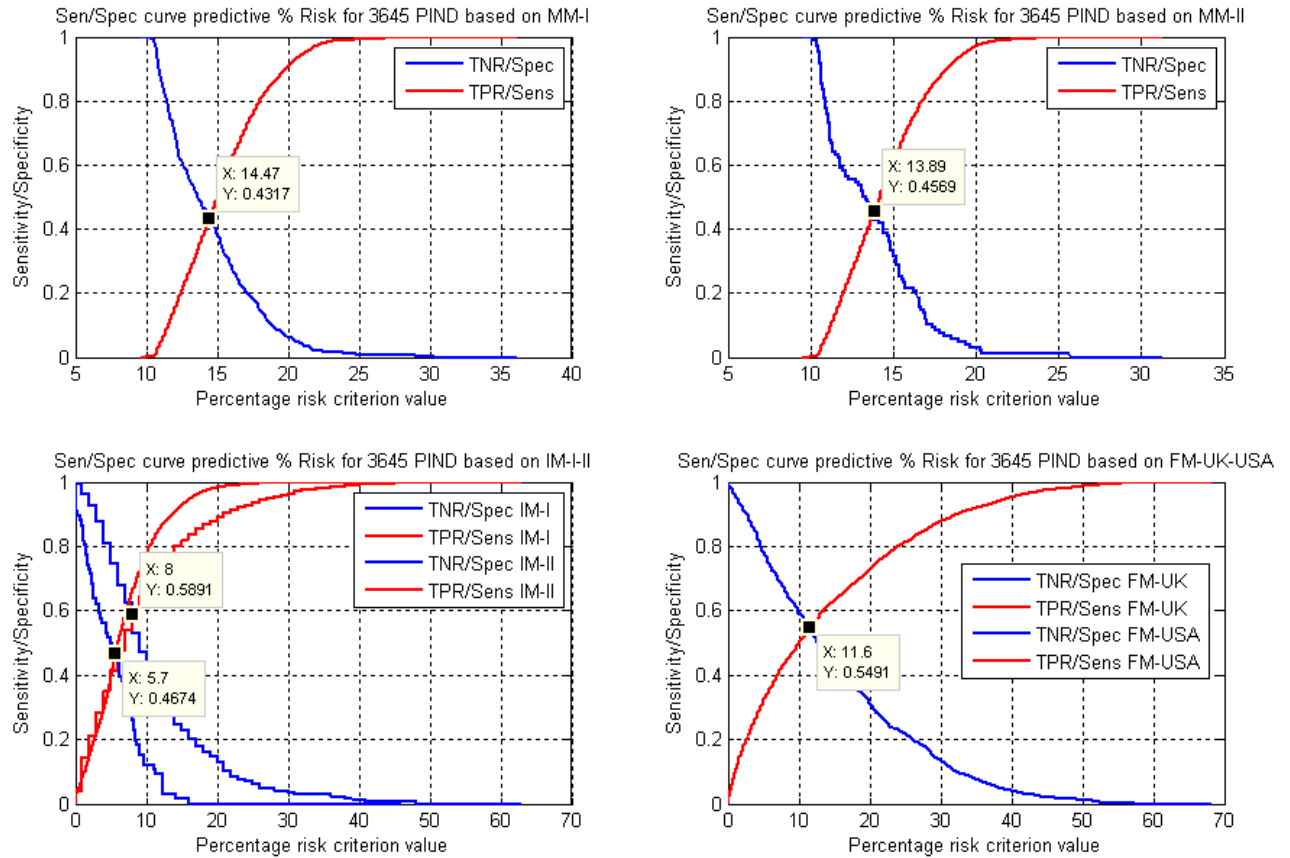


Figure 9.8: Comparison of Sensitivity and Specificity of Predictive Percentage Risk for the 3645 Patient PIND.

In summary, it is inferred from Figure 9.8, that the interception values of the Framingham equations and the proposed CMAUT frameworks are approximately the same. However, it is evident that the Internet Based CVD predictive models give very low interception values as compared to the three other models Figure 9.8. It is difficult to make accurate inference on these values because the discrimination accuracy of Health Informatics systems have not been deeply investigated therefore it is recommended for future research.

9.5.2 Prediction accuracy of the Prediction Models using AUC/ROC:

The Prediction accuracy of Prediction models are determined by calculating the area under curve (AUC) of the model's Receiver Operating Characteristic (ROC). In Chapters 7 and 8, the results of the calculation of the TPR (aka sensitivity) and the FPR (aka specificity) were used to plot the ROC for the determination of the area under curve (AUC).

To determine the AUC, first the TPR (Sensitivity) was plotted on the y-axis and the FPR (specificity) on the x-axis as shown in Figure 9.9.

From the graphs in Figure 9.9, the CMAUT models 1 and 2 curves are located in the positive quadrant which means that the curves are in areas that are higher than the reference diagonal line. When the CMAUT graphs are compared with the Internet CVD calculators, it was observed that the Internet CVD calculator 1 curve lays in the positive quadrant in relation to the reference diagonal line. However, the curve of the Internet CVD calculator 2 lays on the negative quadrant. This denotes that the Internet CVD calculator 1 is more accurate as compared to the Internet CVD calculator 2. Again, the curves for the USA and UK Framingham equations are the same and they both lay in the negative quadrant. This means that the USA and UK Framingham equations are not accurate with respect to the reference diagonal line as indicated in the Figure 9.19.

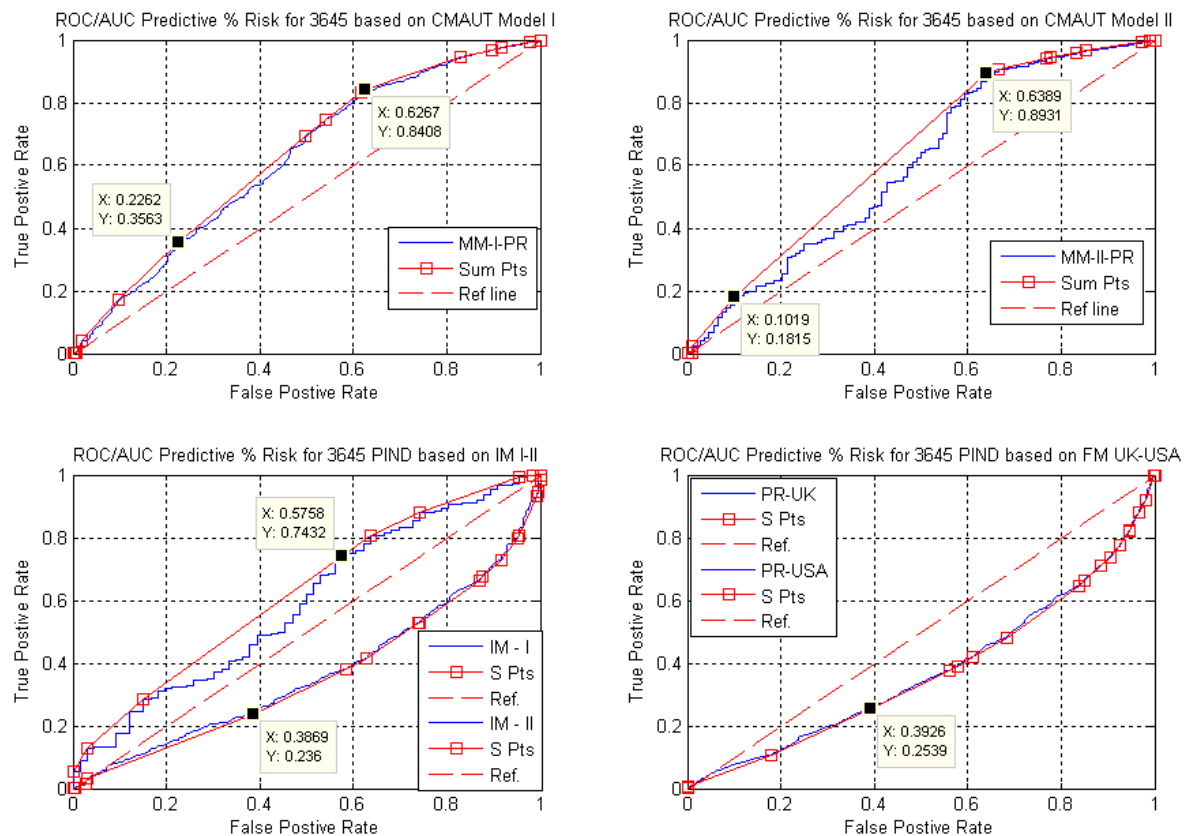


Figure 9.9: Comparison of ROC/AUC of Predictive Percentage Risk values for the 3645 Patient PIND.

- Comparison of the Area under the Curve (AUC) for the Prediction models

For propose of comparison, first the TPR values were plotted against FPR values for each model and value of the area under each curve with reference to the diagonal line calculated. The AUC calculation was done using the Tables 9.10A below that contain the TPR and FPR results of the first 10 participants of the CMAUT model 1 and 2, the two Internet CVD calculators and the Framingham equations. The results of first 30 participants are in Table 9.10B in the Appendix 9, which is the end of the Thesis. The results of the entire 3645 participants are in Table 9.10C in electronic format Appendix 9. The terms specificity and selectivity will be used interchangeably.

The AUC was computed using the Delong Approximate Trapezoidal method in which, the first step is to plot the TPR and FPR curve for each model as indicated in Figure 9.9. Second, the best curve fit method was used to find the equation that best fit each model's curve. This was followed applying the Trapezoidal rule where the area under the curve is split into a number of trapeziums and their areas calculated. Finally the summation technique was used to determine the approximate value of the area under the given curve.

In medical application, area under the curve AUC is computed using the Trapezoidal method as $((Sensitivity\ values + Specificity\ values)/2)$ or $((TPR\ values + FPR\ values)/2)$. In this expression, the diagonal reference line gives a value of 0.5, which is half of the square of the area under consideration. This AUC values for each model was computed using the MATLAB software and the results of the calculation are shown in the Figures in Appendix 9, at the end of this Thesis. Figure 9.10.1 is the CMAUT diagnosis model I from Chapter 5 and Figure 9.10. 2 is the CMAUT diagnosis model II from chapter 5. From the calculations in Figures 9.10.1 and 9.10.2, it is subsumed that CMAUT diagnosis model II has an excellent prediction accuracy of 0.82 while model I has poor prediction ability of 0.22.

Figure 9.10.3 is the CMAUT Prognosis model I from Chapter 7 and Figure 9.10.4 is the CMAUT Prognosis model II from Chapter 7. Inferring from Figures 9.10.3 and 9.10.4, it is subsumed that CMAUT Prognosis model I has an excellent prediction accuracy of 0.92 but model II is just satisfactory prediction ability of 0.55.

Figure 9.10.5 is the Internet CVD model I from Chapter 8 and Figure 9.10.6 is the Internet CVD model II from chapter 8. Therefore from Figures 9.10.5 and 9.10.6, it is inferred that Internet CVD model I has failed as a prediction model because it has a prediction accuracy of 0.08, which is less than 0.5. Again, from Figure 9.9 above the curve of the Internet CVD model I, lays in the negative quadrant but Internet CVD model II is fair prediction since it has prediction ability of 0.675. Similarly, Figure 9.10.7 was computed with the results from UK and USA Framingham equations. From Figure 9.10.7, it is subsumed that Framingham equation models have failed as a prediction model because it has a prediction accuracy of 0.138, which is less than 0.5 and according to Figure 9.9, they lay in the negative quadrant.

In summary, the AUC is used to measure of the discrimination accuracy of prediction models and shows the ability of the models to correctly classify participants with CVD diseases and those without the CVD diseases after performing the test. From the calculations and the curves it is inferred that the CMAUT framework model 1 has a higher prediction value for prognosis but failed in the diagnosis discrimination ability. CMAUT model 2 has better diagnosis discrimination ability but has poor prognosis ability and CMAUT model 2 is the preferred model. This denotes that the CMAUT frameworks can be used as prediction models but they still need to be fine turned to enhance their discrimination ability.

Internet model 2 has better prognosis discrimination ability although this is only fair according to the discrimination guideline Table Viera, (2005). However, Internet model 1 failed because it has a negative value, which means it lays in the negative quadrant and has only -0.088 , which is less than the 0.50 reference value. Similarly, the Framingham equation for both UK and USA, gave a value of -0.1389 , which is failed and lays in the negative quadrant.

These CVD risk calculators are currently been used by people to determine the probability that they are hypertensive or not. According to NICE, (2010), the use of Framingham equation to predict CVD risk in UK population is dangerous because it gives over estimated percentage risk and therefore QRISK must be used. This research has confirmed the recommendation made by NICE, (2010). However, ruling out Framingham equation completely is inadequate and therefore it is recommended further research work should be conducted on this algorithm since the principle is sound but the coefficients may vary from country to country (Zgibor et al., 2006).

Again, a lot of research and benchmarks must be carried on CVD risk calculators before they are uploaded to the websites for public consumption.

9.6 Performance accuracy of Prediction models using Likelihood Ratio.

Likelihood ratio is a technique for evaluating the performance accuracy of diagnosis and prognosis models. Likelihood ratio deals with models with two possible results and handles dichotomous test, where the prediction result must be either YES, which indicates the disease is present or NO when the disease is absent in the participant.

This Likelihood ratio approach is rarely used in medical applications to determine the accuracy of prediction models (Spitalnic, 2004). This is because Likelihood ratio only indicates the interception of the positive Likelihood rate (PLR) and the negative Likelihood rate (NLR) for each participant. This paradigm means Likelihood ratio is a method of testing the level of certainty of the disease in the participant before and after diagnoses or prognosis. This analysis is conducted using the pre-test and post-test probability, which is not covered in this research because the CVD clinical report used has no follow-up information on the participants. However, according to Sanderson et al., (2006), the lower values for negative LR- are acceptable as satisfactory and the higher values for positive LR+ are also acceptable.

In chapter 3, the procedure for determination of Likelihood ratio using the selectivity (aka specificity) and sensitivity approach was outlined as follows:

- Calculate the value of the positive Likelihood ratio $LR+ = (TPR / (1 - TNR))$ and the negative Likelihood ratio $LR- = (1 - TPR) / TNR$ for the entire selected participants in the population.
- Plot all the values of the positive and negative Likelihood ratios on the Y-axis and number of each of the participants PIND on the X-axis as in Figure 9.11.

For comparison reasons, in each graph, the calculated PPR value of each participant was plotted against the participant PIND. See the blue coloured plot in Figure 9.10.

Table 9.10A: Comparison of LRP and LRN for CMAUT models, Internet calculators and Framingham equations for the first 10 participants;

Pserial no.	Grp	LRPM-I)	LRN(M-I)	LRP(M-II)	LRN(M-II)	LRPI-I)	LRN(I-I)	LRP(I-II)	LRN(I-II)	LRP(F-I)	LRN(F-I)
13,956,102.00	No	3484.32	0	3717.472	0	3125	0	3584.22	0	2808.99	0
63,535,102.00	Yes	1745.20	0	1862.197	0	1562.5	0	1788.90	0	1406.47	0
71,831,101.00	No	1162.79	0	1240.695	0	1042.753	0	1761.80	0.01516	1404.7820	0.0012
34,031,101.00	No	872.60	0	930.233	0	781.8608	0	1175.23	0.01516	936.0825	0.0012
72,604,102.00	No	697.84	0	744.048	0	625.3909	0	880.90	0.01516	702.3910	0.0012
31,510,102.00	No	581.73	0	742.7969	0.0017	521.1047	0	704.97	0.01517	561.7548	0.0012
18,633,105.00	No	577.98	0.0065	619.3046	0.0017	446.628	0	587.61	0.01517	468.0412	0.0012
13,008,101.00	Yes	574.23	0.0129	530.7384	0.0017	390.9304	0	503.50	0.01518	401.2857	0.0012
60,417,102.00	No	492.07	0.0129	464.3344	0.0017	347.4635	0	440.64	0.01518	351.0721	0.0012
39,139,101.00	No	430.67	0.0129	463.5530	0.0034	312.6954	0	391.58	0.01519	312.0275	0.0012

9.6.1 Comparison and Interpretation of Likelihood Ratio Graphs

The performance accuracy of prediction models is determined by using the Recalibration and Discrimination techniques discussed in Chapter 3. In chapters 7, the CMAUT prediction models were developed and evaluated using the sensitivity/selectivity, AUC/ROC and Likelihood ratios to determine the performance accuracy of the CMAUT models. In the chapter 8, the PPR value of each participant was determined using the two Internet-based CVD calculators and evaluated with discrimination techniques. Again, in Chapter 8, the PPR value of each participant was calculated using the three Framingham equations and the results evaluated with discrimination techniques.

- Comparison of the CMAUT models 1 and 2 - Observation from the graphs

Figure 9.11 below shows the Likelihood ratios and the PPR values for the CMAUT models 1 and 2. In Figure 9.11, the interception of the positive Likelihood ratio and negative Likelihood ratio curves for the CMAUT model 1 and 2 are on the two participants with PIND number 1600 and 1676 respectively. The Likelihood risk ratio for the participant with PIND number 1600 is 0.997 in model 1 while the participant with PIND 1676 is 1.056 that is, they both have a ratio of approximately one.

According to Spitalnic, (2004), the LR ratio means that the participants with the higher positive values for positive LR+ are acceptable while the model, which have participants with lower negative values for negative LR- are considered as satisfactory. In Figure 9.11, the CMAUT model 1 and 2, all the participants to the left of the point of interception have lower level of certainty of the disease while participants to the right have higher level of uncertainty of the disease. Although there have not been any follow-up on this survey, according to the HSE, (2006) report, few participants who took part in the survey were diagnosed as having hypertension (Craig, et al., 2006b) (Craig, et al., 2008). The Likelihood risk ratios for CMAUT model 1 and 2 are low as compared to the results from the Internet CVD calculators and Framingham equations and therefore confirm the HSE, (2006) findings.

9.6.2 Comparison of the two Internet CVD calculators and Framingham Equations

For the Internet CVD calculators, the interceptions are on participants with PIND numbers 1714 for Internet Model 1 and 2042 for Internet Model 2 while for the Framingham equations the interception is at participant with PIND number 2006. The Likelihood risk ratios are 0.997 and 1.004 for the Internet models 1 and 2 while the Likelihood risk ratio for the Framingham equation is 1.002.

The results are interpreted as follows that for the Internet model 1 fewer participants are on the left hand side of the Internet model 1 interception where the point is (1714; 0.997) as compared to the Internet model 2 interception of (2042; 1.004). This denotes that for Internet model 1 few participants were identified as having hypertension while major of the participants do not have hypertension. The results from Internet model 2 agrees with the results of Framingham equation, that is (2006; 1.002), which indicates that many participants had hypertension. This does not match the HSE, (2006) report analysis, which states that fewer participants were identified as having hypertension during the survey.

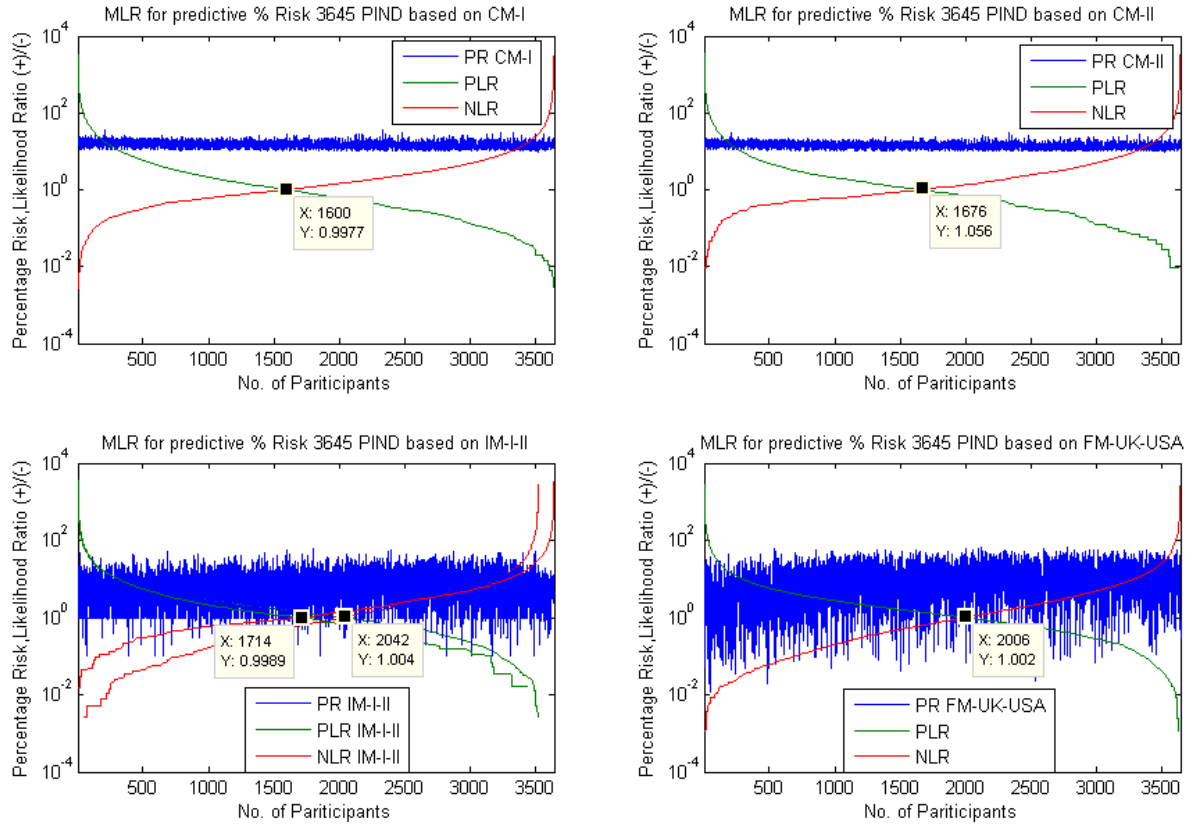


Figure 9.11: Comparison of the Maximum Likelihood Ratio of Predictive Percentage Risk for the 3645 Patient PIND.

From Figure 9.11, it is established that Internet model 2 and Framingham equation results have overestimation PPR values when used in UK population and this is confirmed in the AUC and Likelihood ratio analysis. Internet model 1 and the CMAUT models have good level of performance accuracy in terms of the AUC and Likelihood ratio analysis however more research are required to enhance their calibration and discrimination abilities.

9.7 Summary

From the analysis in this chapter 9, it is subsumed that the hypothesis in this research has been proven with the exception of a few minor issues which are recommended for future research and investigations as discussed in the conclusion Chapter 10. In this chapter, it has been proven that the application of CMAUT Optimisation Framework in CIS reduces the space complexity of clinical data with complementary organs by half the required storage space as compared to non- CMAUT CIS. When substitutable organs are used in the CMAUT Optimisation Framework the result of the space complexity is the same as non- CMAUT CIS. The difference between the data size before and after optimisation with the CMAUT Framework is statistical significant.

The CVD percentage risk values from both the CMAUT Diagnosis and Prognosis Framework model 2 are comparable with NHS Blackheath results but they are different from other existing Risk Prediction models. It was established that the CMAUT model 2 must be enhanced to achieve better prevalence and kappa performance results. It was subsumed that there are inconsistent PPR results from all the CVD risk predictors, therefore more investigation are required.

The prediction accuracy of CMAUT diagnosis model II is excellent with a value of 0.82 while model I is classified as poor prediction model because it has a value of 0.22. However, for Prognosis, the CMAUT model II is just a satisfactory prediction model while the model I has an excellent prediction accuracy of 0.92. Again, from the AUC and Likelihood ratio analysis it was identified that Internet model 2 and the Framingham equations have overestimation PPR values and lay in the negative quadrants. However, Internet model 1 and the CMAUT models have better performance accuracy based on the AUC and Likelihood ratio but their calibration and discrimination abilities need improvement.

Therefore this analysis proves the hypothesis that “Clinical data can be captured with UML class model and re-represented using CMAUT and logical connectors in mathematical format that can be optimised with LP algorithm to reduce the space complexity in CIS and be used as CVD decision tool to predict the risk of users been hypertensive or not”.

Chapter 10: Conclusions and Recommendations for Further Research

10.0 Introduction:

This chapter summarises the overall research and discusses the hypothesis that has been proven as well as the limitations of the proposed CMAUT framework. This is followed by recommendations for future research emanating from the techniques used in this research.

10.1 Conclusions

This research has established that the accumulation of huge amount of data also known as Big Data by data intensive organisations has many benefits. Therefore, data intensive organisations such as Insurance companies, Financial and Medical institutions capture and store gigantic amount of data for their operations and decision making. Social networking websites, such as Facebook and Twitter generate and store the Big Data for marketing and advertisement purposes. Despite these success stories, the storage of data and retrieval of information from these Big Data sets, are associated with problems such as security issues, information highway, interoperability, scalability and information overload.

From the literature reviewed, it was established that, in CIS the issue of information overload has been addressed by the application of clinical data re-representation techniques (Howmotte et al., 1998). For example, EAV/CR uses the data re-representation technique for the design and implementation of CIS. However, according to Nadkarni et al., (2006), experiments conducted prove that EAV/CR has storage and retrieval problems as compared to the relational database. Again, the CISs that use EAV/CR need complex SQL retrieval systems that are not user friendly. The other data re-representation technique used in clinical application is the hybrid application of First Order Logical and Entity Relationship Diagram (FOL/ERD) (De Keizer et al., 2000). The FOL addresses the anomalies and ambiguity issues in CIS but it requires an excellent knowledge of Formal Methods. These models have no optimisation mechanism and therefore they cannot address the problem of information overload in CIS.

This research proposed a new CDSS with UML class model data re-representation and CMAUT framework that optimises the CIS and reduces the space complexity created by information overload. The CMAUT framework can be used as epidemiological tool to predict the percentage risk of a user been hypertensive or not. Based on the proposed solution the hypothesis established that “clinical data can be captured using UML class diagram and re-represented with CMAUT logical connectors to reduce the space complexity in CIS and seamless converted them into the mathematical format that is optimised using LP algorithm to determine the percentage risk of users been hypertensive or not”.

The aim of the proposed CMAUT CVD optimisation framework and the hypothesis in this research were achieved and verified using the following success criteria: - Clinical data was modelled with class diagram and represented using CMAUT that are linked with the AND as well as OR logical connectors. The operation of the CMAUT framework was illustrated using the complementary organs, which are represented using the AND connectors. The UML model shows the classes in the complementary disease domain and the AND logical association between them, which are the heart, kidney and brain. The substitutable organs, kidneys are modelled using super-class and subclasses in the disease domain and their attributes are formalised using the OR logical connectors. The CMAUT statements capture the relation between organs and their attributes in the disease domain using formal method.

The CMAUT expressions were converted into Utility Units using the Utility function and the beta coefficient from the binary logistic regression analysis. The CMAUT expression that uses the total utility units and the attributes in the disease domain, serve as the input to the Optimisation Framework. This analysis confirms the hypothesis that “UML can be used to capture clinical data and the data re-represented with CMAUT technique that can be expressed in mathematical format using AND/ OR logic connectors”. The final mathematical expression was used to formulate the objective function that must be maximised to obtain the optimal value and the comparative variable of each attribute in the combinatory. The objective function was optimised subject to the constraint unit matrix that is built into the algorithm using the MATLAB Linear Programming Software.

The output from the CMAUT framework gives the optimal evaluated value that is converted into Absolute Percentage Risk (APR) and Predictive Percentage Risk (PPR).

The framework gives individual comparative attribute variables that are mapped to each of the participants' input attribute values to determine the diseased attribute in the combinatory. This confirms the hypothesis that "The CMAUT mathematical expression written as objective function can be optimised using LP algorithm subject to the set of inequalities to determine the optimal CVD percentage risk and the deviation of each attribute values from the norm".

From the results of the functional performance of the CMAUT Optimisation Framework, it is concluded that the complementary organs generate less inequality constraints of $O(x + 1)$ as compared to the $O(2x)$ for non-CMAUT system. However, for substitutable organs both CMAUT and non-CMAUT systems generate equal amount of inequality constraints of $y = X^2$, which is expressed as $O(x^2)$ and confirms the hypothesis.

From the statistical analysis, it was established that the data sizes before optimisation was 1216.66 bytes and after optimisation using the CMAUT framework was 463.50 bytes. The T-test gave a p-value of 0.000, which is less than 0.00. This confirms that the difference between the data sizes before and after optimisation are statistically significant and that there is a great reduction in the data size when the CMAUT Optimisation framework is used, which proves the hypothesis.

The APR results from the CVD CMAUT Diagnosis framework models 1 and 2 as well as the PPR results from Prognosis framework models 1 and 2 were benchmarked against CVD Risk Prediction models. The Web based CVD risk calculators used in this research were Internet model 1 from NHS Blackheath and Internet model 2 from Patient UK. The Framingham algorithms used to benchmark the CMAUT framework were the Framingham equation from USA as model I, International equation is model II and the UK equation is model III.

On the operational accuracy of the prediction models, it was established that the kappa and prevalence values for both CMAUT Diagnosis and Prognosis framework model 2 are acceptable. However, the APR and PPR output values from the CMAUT Diagnosis and Prognosis framework model 2 needs further enhancement.

Critical analysis of each of the PPR values for the entire 3456 participants revealed that the results are the same for the Framingham equation model I from USA and the International equation model II but are different in some cases from the UK equation model III. This is because Framingham equation was designed for the US cohorts and not for UK cohorts.

It was identified that the PPR values from the Internet model 1(NHS Blackheath calculator) and the CMAUT Prognosis model 2 are comparable for all the 3564 participants. However the other CVD prediction models reviewed had different PPR values for the same participant hence further research is required.

Analysis of the discriminatory ability of the CVD models revealed that the Internet model II (Patient UK calculator) has better discriminatory accuracy and prediction ability as compared the other models. Internet model II had a decision criterion value of 5.7%, which is very low compared to the 14% from CMAUT models and the NICE recommended value of 20%. To address this limitation, Patient UK has redesigned their website using the QRISK algorithm.

The prediction accuracy of each of the CVD prediction models was determined using ROC and AUC method. It is deduced that the CMAUT diagnosis model II has excellent prediction accuracy but the CMAUT Prognosis model II is just satisfactory as compared to other Prediction models. This means the CMAUT prognosis models can be used as CVD Risk Tool but further work is needed to enhance their prediction accuracy and discrimination ability.

The Performance accuracy of the diagnosis and prognosis prediction models were determined using the Likelihood Risk ratio technique. It was identified that the CMAUT model I has less participants who have lower level of certainty of having the CVD disease as compared to the participants who have higher level of uncertainty of having CVD disease. Therefore, it is deduced that the performance accuracy of the CMAUT models and Internet model 1 (NHS Blackheath) are better than the Internet model 2 and all the three Framingham equations.

Limitations of the CMAUT CVD Framework models I and II are:

- The CMAUT Diagnosis Framework model 2 has reliable APR values and good kappa value, but its prevalence value of 1.03% is classified as poor therefore further improvements are required.

- The CMAUT Prognosis framework model I, gives higher PPR values and hence it is not recommended to be used as epidemiological prediction tool unless further research is conducted on it.
- The discriminatory ability of the CMAUT CVD model 1 is less than 50%; hence it needs to be improved even though it has the highest decision criterion value of 14.5%.
- The prediction accuracy of the CMAUT CVD Framework model 1 was 0.22, which means it failed as Prognosis model 1 therefore further research is required.

10.2 Contribution to Knowledge

This research has addressed the issue of information overload by introducing a novel technique that reduces the space complexity and data size in CIS. Another unique contribution is the development of a new CMAUT decision making model that can be used as epidemiological CVD risk prediction framework. This research makes the following contribution to knowledge:

- A new data re-representation technique where problems in the clinical domain are captured using UML and formulated with CMAUT and logical expressions.
- This new UML – CMAUT technique is an extension to the concept of clinical data re-representation proposed by Haimowitz et al., (1998) and De Keizer et al., (2000).
- This unique CMAUT framework has a new algorithm that converts CMAUT logical expressions into objective functions, which can be optimised using LP technique. The optimal value is seamlessly converted into percentage risk and comparative attribute values, which are presented for medical and clinical decision making.
- This innovative use of CMAUT is unique because the application of combinatorial organs and their multiple attributes to describe diseases has not yet been explored in CIS (Sanderson et al., 2006). Again, a CDSS that uses a hybrid of existing MAUT and combinatorial technique to create CMAUT model is new in medical application:

- A new radical technique in clinical data re-representation using UML class model and CMAUT that can be incorporated with other data representation methods to facilitate mathematical manipulation of clinical and pathological data for health care analysis that is different from other existing methodologies.
- The framework is unique because the in-built algorithm allows the retrieval of optimal amount of data to be transmitted for decision making thus reducing the information overload in CIS as compared to other existing methods.
- This unique partial data retrieval strategy can be applied in areas where Big Data are captured but only partial amount of data are required to be retrieved at any time. This strategy also facilitates the mapping and retrieval of multiple attribute data in combinatorial data intensive systems.
- The framework can be used by data intensive organisations such as financial institutions, and social networking industries to reduce information overload and facilitate data retrieval.
- This new epidemiological risk prediction framework can be used to predict the CVD percentage risk of people who are under 30 years and over 70 years old.
- A new algorithm for the computation of CVD predictive risk based on the Framingham Equation; However, this technique uses measurable CVD attributes to calculate the APR for diagnosis and for the determination of PPR, it uses the arithmetical sum of APR and Predictive time factor $P(T)$, which is based on measurable and non-measurable attributes.

In summary, the proposed new framework will allow medics and medical investigators to receive only the relevant data required to solve problems. This approach localises the clinical data so that only needed data are transferred and thus reduces the amount of data transmitted over the network. It also reduces the amount of information transmitted and stored on handheld devices for any particular disease.

10.3 Further Work and Recommendations

To enhance the predicted output risk values from the diagnosis and predictive frameworks, Software Engineering modelling techniques are recommended for the determination of the accuracy of prediction models. The recommended techniques are the Prequential Likelihood and u –plot methodology (Fenton et al., 1997). Other areas that need to be addressed to improve the output of the medical prediction models are the prediction accuracy, prediction noise and recalibration of the clinical decision support model.

In Software Engineering, the accuracy of prediction models is determined using the Prequential Likelihood function (Fenton et al., 1997). The Prequential Likelihood (PL) method uses three steps to determine the accuracy of prediction models. However, in this research, the Likelihood Ratio method used was recommended for medical application by Sanderson, (2006). This Likelihood Ratio method does not include some of the steps recommended in Software Engineering reliability test. Therefore in future medical prediction models must be benchmarked accordance with the Software Engineering standards.

Another challenge associated with the prediction model is the noisy predicted output values, which fluctuates in magnitude from the true median values. Noise is also known as fluctuation that occurs in prediction models (Fenton et al., 1997). In this research, there are great fluctuations between the calculated percentage predictive risk (PPR) values for the same participants when two different Internet CVD Risk calculators were used, this needs further investigation.

To improve the accuracy of the prediction models, the recalibration u-plot technique must be used to avoid any prejudice in the models' outputs. Different researchers use different methods to calculate and recalibrate prediction model therefore the appropriate technique must be used to calculate the AUC in medical application (Brocklehurst et al., 1990).

- Recommendations for enhancement of CVD Risk Prediction models

There are disagreements between the PPR values from all the CVD Risk Prediction Models, which were examined in this research. Some of the prediction risk calculators gave percentage risk values, which can be interpreted as the participant must die, while in reality they are alive. It is recommended that a detail research must be conduct on the different Web CVD calculators to address the inconsistency in their results. Again, the input risk factors used to design and benchmark CVD Risk Prediction calculators must be standardised.

It was established that the CMAUT diagnosis model 2 has an excellent prediction accuracy of 0.82 but the Prognosis model is satisfactory as compared to all the other Predictive models considered. Therefore the CMAUT models can be used as CVD risk prediction Tool; however, further research must be conducted to fine tune the CMAUT algorithm to enhance the prediction accuracy and discrimination ability of the Framework.

From the AUC and Likelihood ratio analyses conducted, it is concluded that Internet model 2 and the Framingham equations give overestimation PPRs and lay in the negative quadrants. However, Internet model 1 and the CMAUT models have better prediction accuracy for the AUC and Likelihood ratio but their calibration and discrimination abilities need improvement. All the prediction models must be recalibrated to improve their performances.

Finally, from the above discussion it is inferred that the interception values of the Framingham equations and the results from the proposed CVD CMAUT frameworks are comparable. Although, it is evident that the Web Based CVD prediction models give very low interception values as compared to the Framingham and CMAUT models. It is difficult to make an accurate inference on these PPR values because this area of CIS has not been investigated in depth: Therefore it is recommended for future research.

In summary, all the issues and challenges discussed in section 10.3 reveal that Clinical Decision Support Systems and CVD Risk prediction models must be considered as software packages that must meet Software Engineering standards. Therefore, in future all medical and clinical applications must pass reliability test based on Software Engineering principles.

References

- Abdel-Ghaly A. A., Chan, P.Y., and Littlewood, B. (1986). Evaluation of competing software reliability predictions. *IEEE Transactions on Software Engineering*, Vol. 12(9), pp. 950 – 967. doi:10.1109/TSE.1986.6313050
- Abu-Hanna, A., Cornet, R., and de Keizer, N. (2004). The Specification of a Frame-based Medical Terminological System in Protégé. *Studies in health technology and informatics*, Vol. 107, pp. 317.
- Agrawal, D., Das, S., and El Abbadi, A. (2011). Big data and cloud computing: current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology*, ACM, pp. 530-533.
- Anderson, R. J. and Street, P. (1996). Security in Clinical Information Systems. *British Medical Association*. Available from http://httpd.apache.uasw.edu/pub/security/dfn-cert/docs/crypt/Ross_Anderson/policy11.pdf
- Anderson, K. M., Odell, P. M., Wilson, P. W. and Kannel, W. B. (1991). “Cardiovascular disease risk profiles”. *American heart journal*, Vol. 121(1), pp. 293-298.
- Anderson, K. M., Wilson, P., Odell, P. M. and Kannel, W. B. (1991). An updated coronary risk profile: a statement for health professionals. *Circulation*, Vol. 83(1), pp.356-362.
- Anhøj, J. (2003). Generic design of web-based clinical databases. *Journal of Medical Internet Research*, Vol. 5(4).
- Asghari, S. and Mahdavian, M. (2013). Secondary analysis of electronic databases: potentials and limitations. *Diabetologia*, ICD version 10 limitations, pp. 1-2.
- Balaa, F. K., Gamblin, T. C., Tsung, A., Marsh, J. W. and Geller, D. A. (2008). Right hepatic lobectomy using the staple technique in 101 patients. *Journal of Gastrointestinal Surgery*, Vol. 12(2), 338-343.
- Barber, C., Dobkin, D. and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, Vol. 22(4), pp. 469–483. doi:10.1145/235815.235821
- Bath, P. (2008). Health informatics: current issues and challenges. *Journal of Information Science*, Vol. 34(4), pp. 501–518. doi:10.1177/0165551508092267
- Bawden, D., Holtham, C. and Courtney, N. (1999). Perspectives on information overload. In *Aslib Proceedings*, MCB UP Ltd, Vol. 51, No. 8, pp. 249-255.
- Baxt, W. G. (1991). Use of an artificial neural network for the diagnosis of myocardial infarction. *Annals of Internal Medicine*, Vol. 115(11), pp. 843-848.

- Bertot, J. C. and Choi, H. (2013). Big data and e-government: issues, policies, and recommendations. In *Proceedings of the 14th Annual International Conference on Digital Government Research*, ACM, pp. 1-10.
- Beswick, A. and Brindle, P. (2006). Risk scoring in the assessment of cardiovascular risk. *Current opinion in lipidology*, Vol. 17(4), pp. 375-386.
- Bichler, M. and Lee, J. (2001). ABSolute: an intelligent decision making framework for e-sourcing. *Advanced Issues of E-Commerce and Web-Based Information Systems, WECWIS 2001, Third International Workshop*, IBM Thomas J. Watson Res. Center, Yorktown Heights, NY. pp. 195–201, doi:10.1109/WECWIS.2001.933924.
- Billinghurst, M. and Starner, T. (1999). Wearable devices: new ways to manage information. *Computer*, Vol. 32(1), pp. 57-64.
- Bughin, J., Chui, M. and Manyika, J. (2010). Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly*, Vol. 56(1), pp. 75-86.
- Brindle, P., Jonathan, E., Lampe, F., Walker, M., Whincup, P., Fahey, T. and Ebrahim, S. (2003). Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study, *Bmj*, Vol. 327(7426), pp. 1267.
- Brindle, P. M., McConnachie, A., Upton, M. N., Hart, C. L., Smith, G. D. and Watt, G. C. (2005). The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study. *The British Journal of General Practice*, Vol. 55(520), pp. 838.
- Brindle, P., Beswick, A., Fahey, T. and Ebrahim, S. (2006). Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart*, Vol. 92(12), pp. 1752-1759.
- Brownin, D., Loke, S. and Parkinson, B. (2002). A review of the techniques used to build automatic searchable indexes. In: *Proceedings 18th National Conference on manufacturing Research*.
- Campbell, M. J., Machin, D. and Walthers S. J. (2007). *Medical Statistics: A Textbook for the Health Sciences*, Chichester, West Sussex, U.K.:Wiley, 2007, xii+331 pp. 115.
- Cavagna, E., Berletti, R. and Schiavon, F. (2003). Optimized delivery radiological reports: applying Six Sigma methodology to a radiology department. *La Radiologia Medica*, Vol.105(3), pp. 205–14. Available from <http://ukpmc.ac.uk/abstract/MED/12835644>
- Celler, B. G., Lovell, N. H. and Basilakis, J. (2003). Using information technology to improve the management of chronic disease. *Medical Journal of Australia*, 179(5), 242-246.
- Ceriani, R. and Mazzoni, M. (2003). Application of the sequential organ failure assessment score to cardiac surgical patients. *CHEST*, Vol. 123(4), pp. 1229–39. doi:10.1378/chest.123.4.1229

Ceusters, W., Smith, B. and Flanagan, J. (2003). Ontology and medical terminology: *Why description logics are not enough. Towards an Electronic Patient Record (TEPR May 2003)*, Available from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.4053&rep=rep1&type=pdf>

Charatan, Q. and Kans, A. (2004). Formal software development: *From VDM to java*. Basingstoke : Palgrave Macmillan. Available from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Formal+software+development+from+VDM+to+Java#0>

Chen, R.S., Nadkarni, P., Marengo, L., Levin, F., Erdos, J. and Miller, L. (2000). “Exploring performance issues for a clinical database organized using an entity-attribute-value representation”. *Journal of the American Medical Informatics Association*, Vol. 7(5), pp. 475–487. doi:10.1136/jamia.2000.0070475

Chen, H., Chiang, R. H. and Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, Vol. 36(4), pp. 1165-1188.

Choi, B. C. (1998). Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American Journal of Epidemiology*, Vol. 148(11), pp. 1127-1132.

Chuang, J. H., Kukafka, R., Lussier, Y. A., Jenders, R. A. and Cimino, J. J. (2000). A web-based system for prediction of coronary heart disease risk using the Framingham algorithm. *Proc AMIA Symp*, 983. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243802/>

Cimino, J. and Clayton, P. (1994). “Knowledge-based approaches to the maintenance of a large controlled medical terminology”. *Journal of the American Medical Informatics Association*, Vol. 1, pp. 35–50. doi:10.1136/jamia.1994.95236135

Cimino, J. J., Patel, V. L. and Kushniruk, A.W. (2002). The patient clinical information system (PatCIS): technical solutions for and experience with giving patients access to their electronic medical records. *International Journal of Medical Informatics*, Vol. 68(1), pp. 113–127. doi:10.1016/S1386-5056(02)00070-9

Coiera, E. (2003). Guide to Health Informatics, Arnold Publication, Second edition, Hodder Arnold.

Conrick, M. (2006). Health Informatics: Transforming Healthcare with Technology. Thomas Nelson Australia.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, Vol. 115(7), pp. 928-935.

Cornet, R. and Abu-Hanna, A. (2002). Evaluation of a Frame-based Ontology: A Formalization-oriented Approach. *Studies in health technology and informatics*, pp. 488-493.

Cornet, R. and Abu-Hanna, A. (2005). Description logic-based methods for auditing frame-based medical terminological systems. *Artificial Intelligence in Medicine*, Vol. 34(3), pp. 201-217.

Craig, R., Mindell, J. and Hirani, V. (2006a). Health survey for England. Methodology and documentation, pp. 3.

Craig, R., Mindell, J. and Hirani, V. (2006b). Health survey for England. *Obesity and Other Risk Factors in Children*. Available from <http://www.ccsr.ac.uk/esds/events/2012-07-10/craig.pdf>

Craig, R., Mindell, J. and Hirani, V. (2008). Health survey for England 2006. Cardiovascular disease and risk factors in adults. *The Information Centre*, Leeds, pp. 1. Available from <http://www.ccsr.ac.uk/esds/events/2012-07-10/craig.pdf>.

Cui, J. (2009). Overview of risk prediction models in cardiovascular disease research. *Annals of epidemiology, Validation of Diagnosis and prognosis model + AUC*, Vol. 19(10), pp. 711-717.

Cunningham, M. (2009). More than just the kappa coefficient: a program to fully characterize inter-rater reliability between two raters. In *SAS global forum*, pp. 242-2009.

D'Agostino Sr, R. B., Grundy, S., Sullivan, L. M. and Wilson, P. (2001). "Validation of the Framingham coronary heart disease prediction scores". *JAMA: the journal of the American Medical Association*, Vol. **286**(2), pp. 180-187.

D'Agostino, R. B., Russell, M. W., Huse, D. M., Ellison, R. C., Silbershatz, H., Wilson, P. W. and Hartz, S. C. (2000). "Primary and subsequent coronary risk appraisal: new results from the Framingham study". *American heart journal*, Vol. **139**(2), pp. 272-281.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M. and Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care the Framingham Heart Study. *Circulation*, Vol. 117(6), pp. 743-753.

De Keizer, N., Abu-Hanna, F. A. and Abu-Hanna, A. (2000). Understanding Terminological Systems II: Experience with Conceptual and Formal Representation of Structure. *Methods of Information in Medicine*, Vol. 39(1), 22–29. Available from <http://www.schattauer.de/de/magazine/uebersicht/zeitschriften-a-z/methods/contents/archive/issue/special/manuscript/56/download.html>

De Keizer, N., Abu-Hanna, F. A. and Abu-Hanna, A. (2000). Understanding terminological systems I: terminology and typology. *Methods of Information in Medicine*, Vol. 39(1), pp. 16–21. Available from <http://www.schattauer.de/en/magazine/subject-areas/journals-a-z/methods/contents/archive/issue/708/manuscript/55/download.html>

De Keizer, N. F., Abu-Hanna, A., Cornet, R., Zwetsloot-Schonk, J. H. M. and Stoutenbeek, C. P. (1999). Analysis and design of an ontology for intensive care diagnoses. *Methods of information in Medicine*, Vol. 38, pp.102-112.

De Mendonça, A., Vincent, J. L., Suter, P. M., Moreno, R., Dearden, N. M., Antonelli, M. and Cantraine, F. (2000). Acute renal failure in the ICU: risk factors and outcome evaluated by the SOFA score. *Intensive care medicine*, Vol. 26(7), pp. 915-921.

- Denton, D. K. and Richardson, P. (2012). “Using intranets to reduce information overload”. *Journal of Strategic Innovation and Sustainability*, Vol. 7(3), pp. 84-94.
- Denley, I. and Smith, S. (1998).” Implementing access control to protect the confidentiality of patient information in clinical information systems in the acute hospital”. *Health Informatics Journal*, Vol. 4(3-4), pp. 174–178. doi:10.1177/146045829800400307
- Driscoll, A. O., Daugelaite, J. and Sleator, R. D. (2013). “‘Big Data’, Hadoop and Cloud Computing in Genomics”. *Journal of biomedical informatics*. Vol. 8(6), pp 100-110
- Deutsch, T., Carson, E. and Ludwig, E. (1994). Dealing with medical knowledge: Computers in Clinical Decision Making. Plenum Press. Available from <http://dl.acm.org/citation.cfm?id=562069>
- Donnan, P. T., Donnelly, L., New, J. P. and Morris, A. D. (2006). Derivation and validation of a prediction score for major coronary heart disease events in a UK type 2 diabetic population. *Diabetes Care*, Vol. 29(6), pp. 1231-1236.
- Edoh, A. A. (2004). Combinatorial Multi-Objective Auction (CMOA) A new form of automatic auction protocol. MPhil Thesis in Computing Science, City University, London, UK.
- Edoh, A. A. (2005). Combinatorial Multi-attribute Auction (CMOA) Framework for E-Auction. *In the proceedings of IEEE/SCSS conference on Advances in Computer, Information and Systems Sciences, and Engineering*. Springer.
- Edoh, A., Imafidon, C., Kans, A. and Thiyagu, R. (2011). A New Framework for Optimising Clinical Information Systems (CIS). *Adv. Comp and Tech*, Vol. 2, pp. 195–206.
- Falaschetti, E., Chaudhury, M., Mindell, J. and Poulter, N. (2009). Continued improvement in hypertension management in England: results from the Health Survey for England 2006. *Hypertension*, Vol. 53(3), pp. 480–486. doi:10.1161/HYPERTENSIONAHA.108.125617
- Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in medicine*, Vol. 21(20), pp. 3093-3106.
- Faughnan, J. (1997). Evaluation Methods in Medical Informatics. *BMJ*. Available from <http://www.bmj.com/content/315/7109/689.1?variant=full>
- Fenton, N. E. and Pledger, S. L. (1997). 'Software Metrics: A Rigorous and Practical Approach', “The SERENE Method Manual EC Project No. Software Measurement, Atlanta, USA, October, pp. 157-224.
- Fenton, N. E. and Neil, M. (2000). Software metrics: roadmap. In *Proceedings of the Conference on the Future of Software Engineering*, ACM, pp. 357-370.
- Fernandes, L., O'Connor, M. and Weaver, V. (2012). “Big data, bigger outcomes: Healthcare is embracing the big data movement, hoping to revolutionize HIM by distilling vast collection of data for specific analysis”. *Journal of AHIMA/American Health Information Management Association*, Vol. 83(10), pp. 38-43.

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology*, pp. 1-3.

Friedman, C., Liu, H. and Shagina, L. (2001). Evaluating the UMLS as a source of lexical knowledge for medical language processing. *Proc AMIA Symp*, pp. 189–93. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243298/>

Friedman, C. and Hripcsak, G. (1990). A generalized relational schema for an integrated clinical patient database. *Proc Annu Symp Comput Appl Med Care*, Vol. 7, pp. 335–339. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245527/>

Funk, B., Möller, A. and Niemeyer, P. (2009). A reference architecture for the integration of EMIS and ERP-systems. *Lecture Notes in Informatics*, Vol. 154, pp. 3393-3401.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of computational biology*, Vol. 7(3-4), pp. 601-620.

Graaf, K., Van, d., Rhees, R. and Palmer, S. (1997). *Schaum's Outline on Human Anatomy and Physiology*. McGraw-Hill. Available from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Schaum's+Outline+of+Human+Anatomy+and+Physiology#1>

Granger, R. (2006). In the 15th International World Wide Web Conference. Retrieved 21 June 2011.

Green, A. (2011). “Danish clinical databases: An overview”. *Scandinavian journal of public health*, Vol. 39(7), pp. 68-71.

Gjertsen, F., Bruzzzone, S., Vollrath, M. E., Pace, M. and Ekeberg, O. (2013). Comparing ICD-9 and ICD-10: The impact on intentional and unintentional injury mortality statistics in Italy and Norway. *Injury, Introduction of ICD version 11 and 12*, Vol. 44(1), pp 132-138.

Grunkemeier, G. L. and Jin, R. (2001). Receiver operating characteristic curve analysis of clinical risk models. *The Annals of thoracic surgery*, Vol. 72(2), pp. 323-326.

Goulding, A. (2001). “Information poverty or overload?”. *Journal of Librarianship and Information Science*, Vol. 33(3), pp. 109-111.

Guzder, R. N., Gatling, W., Mullee, M. A., Mehta, R. L. and Byrne, C. D. (2005). Prognostic value of the Framingham cardiovascular risk equation and the UKPDS risk engine for coronary heart disease in newly diagnosed type 2 diabetes: results from a United Kingdom study. *Diabetic Medicine*, Vol. 22(5), pp. 554-562.

Guyton, A. and Hall, J. (1992). *Human physiology and mechanisms of disease*. Available from <http://www.lavoisier.fr/livre/notice.asp?id=OL3WRXA26XROWJ>

Hackney, R. and Dhillon, G. (1998). Developing a Clinical Information System: an interpretive case analysis within a UK hospital. *Health Serv Manage Res*, Vol. 11(4), pp. 238–245. Available from <http://www.ncbi.nlm.nih.gov/pubmed/10338692>

- Haimowits, I., Patil, R. and Szolovits, P. (1988). Representing medical knowledge in a terminological language is difficult. *Application in Medical Care*, Vol. 9, pp. 101–109. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245250/>
- Hellman, R. (2001). Improving patient safety by reducing medical errors: sleuthing strategies for the endocrinologist. A workshop presented at the American Association of Clinical Endocrinology 10th Scientific Sessions, San Antonio, Tex., May 3.
- Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F. B. and Babu, S. (2011). Starfish: A Self-tuning System for Big Data Analytics. In *CIDR*, Vol. 11, pp. 261–272.
- Herodotou, H. and Babu, S. (2011). Profiling, what-if analysis, and cost-based optimization of MapReduce programs. *Proc. of the VLDB Endowment*, Vol. 4(11), pp. 1111–1122.
- Higgins, B., Williams, B., Bakhshi, L. and Brown, M. (2006). Management of hypertension in adults in primary care: partial update. Royal College of Physicians, Aug 18, pp. 103–139. Available from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Management+of+hypertension+in+adults+in+primary+care:+partial+update#0>
- Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M. and Brindle, P. (2007). “Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study”. *BMJ: British Medical Journal*, Vol. 335(7611), pp. 136.
- Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A. and Brindle, P. (2008). Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, Vol. 336(7659), pp. 1475–1482.
- Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M. and Brindle, P. (2007). “Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study”. *BMJ: British Medical Journal*, Vol. 335(7611), pp. 136.
- Hoelzer, S., Schweiger, R. and Dudeck, J. (2003). Knowledge representation in digital medical documents for efficient information access and retrieval. *Studies in Health Technology Inform*, Vol. 95, pp. 451–456. Available from <http://www.ncbi.nlm.nih.gov/pubmed/14664028>
- Hoppszallern, S. (2003). Health care benchmarking 2003. Hospitals & health networks/AHA, Vol. 77(2), pp. 37–44. Available from <http://www.ncbi.nlm.nih.gov/pubmed/12633068>
- Ho, J. and Tang, R. (2001). Towards an optimal resolution to information overload: an infomediary approach. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, pp. 91–96.
- Huff, S. and Berthelsen, C. (1991). Evaluation of an SQL model of the HELP patient database. *Proc Annu Symp Comput Appl Med Care*, IEEE Computer Press, Los Alamitos, CA, Washington, DC, pp. 386–390. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2247560/>

Hunter, L. (1998). Portfolio 2000: Managing Clinical Systems. Clin Lab Manage Rev, Vol. 12(5), pp. 305–309. Available from <http://ukpmc.ac.uk/abstract/MED/10185008>

Hsieh, S. H., Hsieh, S. L., Cheng, P. H. and Lai, F. (2012). E-health and healthcare enterprise information system leveraging service-oriented architecture. Telemedicine and e-Health, Vol. 18(3), pp. 205-212.

HSE. (2006). Survey REPORT SN: 5809 National Centre for Social Research and University College London. Department of Epidemiology and Public Health, Health Survey for England, 2006. [computer file]. 4th Edition. Colchester, Essex: UK Data Archive [distributor], July 2011. SN: 5809, Available from <http://dx.doi.org/10.5255/UKDA-SN-5809-1>; <http://discover.ukdataservice.ac.uk/catalogue/?sn=5809&type=Data%20catalogue>

HSE. (2006). User Guide Health Survey for England 2006; Cardiovascular and risk factors UK Data Archive Study Number 5809, Available from <https://catalogue.ic.nhs.uk/publications/public-health/surveys/heal-surv-cvd-risk-obes-ad-ch-eng-2006/heal-surv-cvd-risk-obes-ad-ch-eng-2006-rep-v3.pdf>

HSE. (2008). REPORT User Guide . UK Data Archive Study Number 6397, Health Survey for England: Available from www.esds.ac.uk/doc/6397/mrdoc/pdf/6397userguide.pdf

Hripcsak, G., Clayton, P. D., Pryor, T. A., Haug, P., Wigertz, O. B. and Van der Lei, J. (1990). The Arden Syntax for Medical Logic Modules. In Proceedings/the... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care, pp. 200-204.

Iakovidis, I., Wilson, P. and Healy, J. (2004). *E-health: current situation and examples of implemented and beneficial e-health applications*. IOS Press. Available from <http://books.google.com/books?hl=en&lr=&id=oEE-45obxQ8C&oi=fnd&pg=PR5&dq=E-Health:Current+Situation+and+example+of+implemented+and+Beneficial+E-health+Application&ots=gs0Z1EeLgy&sig=sX1oz6Bc9scLwQK-BryOPJkcEfA>

Imafidon, C., Kans, A. and Edoh, A. A. (2009). Human organ re-representation using UML and CMAUT. The School of Computing and Technology 4th Annual Conference, University of East London , ICGES Press, pp. 235–245. Available from <http://dspace.uel.ac.uk/jspui/handle/10552/927>

IBM Report (2013). "IBM What is big data? — Bringing big data to the enterprise". Available from <http://www.ibm.com>.

Lakhani, A., Coles, J., Eayres, D., Spence, C. and Rachet, B. (2005). Information in practice Creative use of existing clinical and health outcomes data performance indicators closely linked to clinical care. BMJ, Vol. 330(June), pp. 330–1426.

Laudon, K. C. and Laudon, J. P. (2011). *Essentials of management information systems*, Boston: Prentice Hall.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S. and Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, Vol. 52(2), pp. 21-31.

Lisboa, P. J. and Taktak, A. F. (2006). The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, Vol. 19(4), pp. 408-415.

Lloyd-Jones, D. M. (2010). Cardiovascular Risk Prediction Basic Concepts, Current Status, and Future Directions. *Circulation*, Vol. 121(15), pp. 1768-1777.

Lopez, D. M. and Blobel, B. G. (2009). “A development framework for semantically interoperable health information systems”. *International journal of medical informatics*, Vol. 78(2), pp. 83-103.

Jacobson, N. S. and Truax, P. (1991). “Clinical significance: a statistical approach to defining meaningful change in psychotherapy research”. *Journal of consulting and clinical psychology*, Vol. 59(1), pp. 12.

Jay, M. T. (2006). *Super Review Anatomy and Physiology by the Staff of Research & Education Association*.

Jenders, R. A. and Cimino, J. J. (2000). A web-based system for prediction of coronary heart disease risk using the Framingham algorithm. *Proc AMIA Symp*, (February), Vol. 983. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243802>

Johnson, S. (1996). “Generic Data Modeling for Clinical Repositories”. *Journal of the American Medical Informatics Association*, Vol. 3(5), pp. 328–39. doi:10.1136/jamia.1996.97035024

Johnson, S. B., Terre, P. D., Anna, P., Hospital, P. and York, N. (1997). Generic Database Design for Patient Management Information. *Proc AMIA Annu Fall Symp*, pp. 22–26. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233478/>

Johnson, S. and Chatziantoniou, D. (1999). Extended SQL for manipulating clinical warehouse data. *Proc AMIA Symp*, pp. 819–823. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232585/>

Krause, T., Lovibond, K., Caulfield, M., McCormack, T. and Williams, B. (2011). Management of hypertension: summary of NICE guidance. *BMJ*, pp. 343.

Kremelberg, D. (2011). *Practical Statistics: A Quick and Easy Guide to IBM SPSS statistics, STATA, and other Statistical Software*. SAGE Publications, Inc.

Manchikanti, L., Falco, F. J. and Hirsch, J. A. (2013). “Ready or not! Here comes ICD-10”. *Journal of neurointerventional surgery*, Vol. 5(1), pp. 86-91.

Massimo, F. F. (1998). The standard ‘Healthcare Information Systems Architecture and the DHE middleware. *International Journal of Medical Informatics*, 52(1), 39-51.

Mattmann, C. (2003). *A software Architecture based Framework for highly Distribute and data intensive scientific application*.

May, M., Lawlor, D. A., Brindle, P., Patel, R. and Ebrahim, S. (2006). Cardiovascular disease risk assessment in older women: can we improve on Framingham? British Women's Heart and Health prospective cohort study. *Heart*, Vol. 92(10), pp. 1396-1401.

Moreno, R., Vincent, J. L., Matos, R., Mendonca, A., Cantraine, F., Thijs, L. and Willatts, S. (1999). The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study. *Intensive care medicine*, Vol. 25(7), pp. 686-696.

Nadkarni, P. (2002). *An introduction to entity-attribute-value design for generic clinical study data management systems*. Available from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:An+introduction+to+entity+attribute+value+design+for+generic+clinical+study+data+management+systems#0>

Nadkarni, P. M. and Brandt, C. (1998). "Data extraction and ad hoc query of an entity-attribute-value database". *Journal of the American Medical Informatics Association*, Vol. 5(6), pp. 511-27. Available from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=61332&tool=pmcentrez&rendertype=abstract>

Nadkarni, P. M., Marenco, L., Chen, R., Skoufos, E., Shepherd, G. and Miller, P. (1999). "Organization of heterogeneous scientific data using the EAV/CR representation". *Journal of the American Medical Informatics Association*, Vol. 6(6), pp. 478-493. doi:10.1136/jamia.1999.0060478

Nadkarni, P. M. and Brandt, C. (1998). Managing attribute-value clinical trials data using the ACT/DB client-server database system. *Journal of the American Medical Informatics Association*, Vol. 5(2), pp. 139-51. doi:10.1136/jamia.1998.0050139

Nadkarni, P. M., Brandt, C. M. and Marenco, L. (2000). "WebEAV Automatic Metadata-driven Generation of Web Interfaces to Entity-Attribute-Value Databases". *Journal of the American Medical Informatics Association*, Vol. 7(4), pp. 343-356. doi:10.1136/jamia.2000.0070343

Nainil, C. and Chheda, M. (2005). Electronic Medical Records and Continuity of Care Records-The Utility Theory. *Application of Information Technology and Economics*. Available from [#0](http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Electronic+Medical+Records+and+Continuity+of+Care+Records+-+The+Utility+Theory)

NICE (2006). Hypertension - About NICE guidance; Hypertension: Quick Reference Guide. The published NICE clinical guideline, Available from <http://guidance.nice.org.uk/CG34>

NICE (2010). Prevention of cardiovascular disease - About NICE guidance PH25 - Issued: June 2010. The guidance for development and running of effective regional CVD prevention programmes Quick Reference Guide. Available from <http://guidance.nice.org.uk/PH25>

Olsen, K., Sochats, K. and Williams, J. (1998). Full text searching and information overload. *The International Information & Library Review*, Vol. 30(2), pp.105-122. doi:10.1006/iilr.1998.0087

- O'Carroll. and Patrick, W. (2003). *Public health informatics and information systems*. Springer. Available from <http://books.google.com/books?hl=en&lr=&id=ap6uCR0Ybo4C&oi=fnd&pg=PR5&dq=Public+Health+Informatics+and+Information+Systems&ots=B6RGE094KE&sig=mB1tNDHMBbpufSANuAwuuQi68RY>
- Odell, P. M., Anderson, K. M. and Kannel, W. B. (1994). New models for predicting cardiovascular events. *Journal of clinical epidemiology*, Vol. 47(6), pp. 583-592.
- Page, T. (2012). "Healthcare Innovations: Service and Product Design", Lambert Academic Publishing, ISBN: 978-3-659-12465-5.
- Park, H., Yoo, S., Kim, B., Choi, J. and Chun, J. (2003). Optimizing Query Response with XML User Profile in Mobile Clinical Systems. *AMIA Annual Symposium*, 963. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480119/>
- Panagiotakos, D. B., and Stavrinou, V. (2006). Methodological issues in cardiovascular epidemiology: the risk of determining absolute risk through statistical models. *Vascular health and risk management*, Vol. 2(3), pp. 309.
- Paul, R. and Hoque, A. S. M. L. (2011). Optimized Entity Attribute Value Model: A Search Efficient Representation of High Dimensional and Sparse Data. *Interdisciplinary Bio Central*, Vol. 3(1), pp. 9.
- Petrounias, I. and Kodogiannis, V. (2006). A software engineering framework for biomedical diagnostic systems. *Proceedings of the 2006 international workshop on Workshop on interdisciplinary software engineering research*, pp. 61 – 64. doi:10.1145/1137661.1137675
- Pencina, M. J., D'Agostino, R. B., Larson, M. G., Massaro, J. M. and Vasan, R. S. (2009). Predicting the 30-year risk of cardiovascular disease The Framingham Heart Study. *Circulation*, Vol. 119(24), pp. 3078-3084.
- Poppendieck, M. and Poppendieck, T. (2003). *Lean Software Development: An Agile Toolkit for Software Development Managers*. Addison-Wesley Professional.
- Raman, R. and Grossmann, I. (1994). Modelling and computational techniques for logic based integer programming. *Computers & Chemical Engineering*, Vol. 18(7), pp. 563–578. doi:10.1016/0098-1354(93)E0010-7
- Rassinoux, A. and Miller, R. (1998). Modeling Concepts in Medicine for Medical Language Understanding. *Methods of Information in Medicine*, Vol. 37(4-5), pp. 361–72. Available from <http://ukpmc.ac.uk/abstract/MED/9865034>
- Raymond, B. and Dold, C. (2001). Clinical information systems: achieving the vision. *The Benefits of Clinical Information Systems*. Available from http://web.uvic.ca/~h351/hinf351_course_data/Raymond_CIS-Achieving_the_vision.pdf
- Rodrigues, J. (2009). *Health information systems: concepts, methodologies, tools and applications*. IGI Global.

- Russom, P. (2011). Big data analytics. *TDWI Best Practices Report, Fourth Quarter*.
- Saaty, T. L. (2003). Decision-making with the AHP: Why is the principal eigenvector necessary. *European journal of operational research*, Vol. 145(1), pp. 85-91.
- Saaty, T. L. and Vargas, L. G. (2000). Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process. *The Analytic Hierarchy Process*, Vol. 6.
- Safran, C. and Chute, C. (1995). “Exploration and exploitation of clinical databases”. *International Journal of Bio-Medical Computing*, Vol. 39(1), pp. 151–156. doi:10.1016/0020-7101(94)01094-H
- Sanderson, C. G. R. (2006a). *Analytical Models for Decision Making Understanding Public Health*. Open University Press; Pap/Cdr edition.
- Sherlaw-Johnson, C. and Mitchard, J. (1995). Displaying the long-term progression of patients with coronary artery disease. *British Heart Journal*, Vol. 74(5), pp. 559–62. doi:10.1136/hrt.74.5.559
- Sheridan, S., Pignone, M. and Mulrow, C. (2003). “Framingham-based tools to calculate the global risk of coronary heart disease”. *Journal of general internal medicine*, Vol. 18(12), pp. 1039-1052;
- Spronk, R. (2007). The Spine, an English National Programme for NHS. Available from http://www.ringholm.de/docs/00970_en.htm Retrieved 20 -08-2011
- Silberschatz, K. (2001). *Database system concepts with Oracle CD*. Columbus, OH: McGraw-Hill Science/Engineering/Math. Available from <http://www.lavoisier.fr/livre/notice.asp?ouvrage=1033557>
- Smith, C. A. (2006). Information retrieval in medicine: The electronic medical record as a new domain, A. Grove, Ed. *Proceedings 69th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, Vol. 43(1), pp. 1–30. doi:10.1002/meet.1450430190
- Steve, R. (2006). *The nine projects at the heart of NHS IT Public Sector Review*. Silicon.com. Available from <http://www.silicon.com/publicsector/>.
- Stevens, R. J., Kothari, V., Adler, A. I., Stratton, I. M. and Holman, R. R. (2001). The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clinical Science*, Vol. 101(6), pp. 671-679.
- Sujansky, W. and Altman, R. (1994). Towards a standard query model for sharing decision-support applications. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association, pp. 325.
- Templin J. M. (2006). *Super Review Anatomy and Physiology by the Staff of Research & Education Association*.

Taylor Paul. (2006). From Patient Data to Medical Knowledge: *Principles and Practice of Health Informatics*. Oxford: Blackwell Publishing Ltd.

Tveito, A. and Hasvold, P. (2002). Requirements in the medical domain: Experiences and prescriptions. *Software, IEEE*, Vol. 19(6), pp. 66 – 69. doi:10.1109/MS.2002.1049394

Tzellepis, T. Z. (2004). *Design and implementation of two level clinical information systems (CIS) based on Archetypes*.

Van Velsen, L., Beaujean, D. J. and van Gemert-Pijnen, J. E. (2013). Why mobile health app overload drives us crazy, and how to restore the sanity. *BMC Med. Inf. & Decision Making*, Vol. 13, pp. 23.

Velde, R. Van, D. (2000). “Framework for a clinical information system”. *International Journal of Medical Informatics*, Vol. 57(1), pp. 57–72. doi:10.1016/S1386-5056(99)00062-3

Viera, A. J. and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, Vol. 37(5), pp. 360-363.

Vincent, J. L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H. and Thijs, L. G. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive care medicine*, Vol. 22(7), pp. 707-710.

Wang, P., Pryor, A., Narus, S., Hardman, R. and Deavila, M. (1997). The Web Enabled IHC Enterprise Data Warehouse for Clinical Process Improvement and Outcomes Measurement. *Proc American Medical Informatics Association Annu Fall Symp*, 1028., 8280.

Williams, B., Poulter, N. R., Brown, M. J., Davis, M. and McInnes, G. T.P. J. (2004). British Hypertension Society guidelines for hypertension management 2004 (BHS-IV): summary. *BMJ*, Vol. 328, pp. 634–640. doi:10.1136/bmj.328.7440.634

Woodward, M., Brindle, P. and Tunstall-Pedoe, H. (2007). Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart*, Vol. 93(2), pp. 172-176.

Wilson, P. W., D’Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H. and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, Vol. 97(18), pp. 1837-1847.

Wolf, P. A., D’Agostino, R. B., Belanger, A. J. and Kannel, W. B. (1991). Probability of stroke: a risk profile from the Framingham Study. *Stroke*, Vol. 22(3), pp. 312-318.

Williams, B., Poulter, N. R., Brown, M. J., Davis, M., McInnes, G. T., Potter, J. F. and Thom, S. M. (2004). “Guidelines for management of hypertension:” Report of the fourth working party of the British Hypertension Society, 2004-BHS IV. *Journal of human hypertension*, Vol. 18(3), pp 139-185.

Winsten, D. and Carroll, R. (1996). Optimizing clinical information systems in complex computing environments. Panel discussion. *Towards an Electronic Patient Record (TEPR 2003)*, San Antonio, May 2003, Boston, Vol.10(4), pp. 47–58.

Yusuf, H. R., Giles, W. H., Croft, J. B., Anda, R. F. and Casper, M. L. (1998). Impact of multiple risk factor profiles on determining cardiovascular disease risk. *Preventive medicine*, Vol. 27(1), pp. 1-9.

Yu, C. S. (2002). A GP-AHP method for solving group decision-making fuzzy AHP problems. *Computers & Operations Research*, Vol. 29(14), pp. 1969-2001.

Yoo, S., Kim, B., Park, H. and Choi, J. C. J. (2003). Realization of real-time clinical data integration using advanced database technology. *AMIA Annual Symposium*, pp. 738–742. Available from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479960/>

Zgibor, J. C., Piatt, G. A., Ruppert, K., Orchard, T. J. and Roberts, M. S. (2006). Deficiencies of cardiovascular risk prediction models for type 1 diabetes. *Diabetes Care*, Vol. 29(8), pp. 1860-1865.

Zikopoulos, P., Deroos, D., Parasuraman, K., Deutsch, T., Corrigan, D. and Giles, J. (2012). Harness the Power of Big Data: The IBM Big Data Platform, Available from <http://www.ibmbigdatahub.com/whitepapers>

Zhang, Y. (1995). Solving Large-Scale Linear Programs by Interior-Point Methods Under the MATLAB Environment. Technical Report TR96-01. doi:10.1080/10556789808805699

Appendix

Appendix 3.0 for Chapter 3

The following appendices are in electronic format on the attached USB

1. Appredix 3.1C: hse06ai_Full data with 21299 records
2. Appredix 3.2C: hse06ai_sample1_refined_ with 9194 records
3. Appredix 3.3C: CVD data set for over 16 years old Model I_ 4316 records
4. Appredix 3.4C: CVD data set for over 30 years old Model 2_ 3645 records

Appendix 5.0 for Chapter 5

Table 5.8B: Raw data of the first 30 participants used in Chapter 5:

Serial no.	Grp	BpI	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HD L	MAP	DIA	TC	SMK	CV D	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
71,831,101.00	No	No	66	Women	White	89.00	14.32	159.00	70.00	1.90	99.50	No	6.90	No	No	Yes
34,031,101.00	No	No	84	Women	White	48.50	16.17	112.00	63.50	2.20	80.00	No	5.00	Yes	Yes	Yes
72,604,102.00	No	No	59	Women	White	36.50	16.19	109.50	73.00	2.00	85.00	No	6.00	No	No	No
13,008,101.00	Yes	Yes	50	Women	White	43.00	16.65	117.00	74.00	1.70	88.50	No	6.00	Yes	Yes	Yes
39,139,101.00	No	No	34	Women	White	48.00	16.81	102.00	54.00	1.80	70.00	No	6.50	Yes	No	No
47,856,102.00	No	No	51	Women	White	44.00	16.85	100.50	56.50	1.90	71.00	No	5.10	No	No	No
37,710,101.00	No	No	61	Women	White	43.00	17.43	120.00	77.00	1.20	91.50	No	5.50	No	No	No
54,256,101.00	No	No	31	Women	White	40.00	17.72	124.00	84.00	2.00	97.00	No	3.90	Yes	No	No
53,817,101.00	No	No	43	Women	White	39.00	17.72	107.00	68.00	1.70	81.00	No	4.90	Yes	No	No
44,633,102.00	No	No	39	Women	White	39.00	17.87	106.00	67.00	1.40	80.00	No	5.30	No	No	No
34,523,102.00	No	No	45	Women	White	51.50	18.09	120.00	68.50	2.00	85.50	No	4.70	No	No	No
72,323,102.00	No	No	38	Women	White	39.00	18.10	116.00	77.00	1.40	90.00	No	4.20	Yes	Yes	No
42,831,101.00	No	No	35	Men	White	64.50	18.15	135.00	70.50	1.70	92.00	No	4.10	Yes	No	No
72,833,102.00	No	No	40	Men	White	58.50	18.19	117.00	58.50	1.50	78.00	No	4.60	Yes	No	No
15,047,101.00	No	No	86	Women	White	79.00	18.27	138.00	59.00	2.50	85.50	No	8.90	Yes	No	Yes
23,202,102.00	No	No	58	Women	White	48.00	18.29	113.50	65.50	1.10	81.50	No	3.60	Yes	No	No
20,408,101.00	No	No	44	Women	White	38.00	18.41	100.00	62.00	0.90	75.00	No	4.90	No	No	No
76,910,102.00	No	No	48	Women	White	46.00	18.43	110.00	64.00	2.00	79.00	No	8.00	Yes	No	No
13,856,101.00	No	No	51	Women	White	32.50	18.45	92.50	60.00	2.60	70.50	No	6.70	Yes	No	Yes
21,413,101.00	Yes	Yes	86	Men	White	65.50	18.48	149.00	83.50	1.20	105.00	No	5.60	No	Yes	No
25,713,101.00	No	No	38	Women	White	32.50	18.55	103.50	71.00	1.30	81.50	No	4.00	Yes	No	No
29,646,101.00	No	No	32	Women	White	36.00	18.58	113.50	77.50	1.40	89.50	No	4.60	No	No	No
59,015,101.00	No	No	31	Women	Asian or Asian British	38.00	18.59	100.00	62.00	1.50	74.50	No	4.50	Yes	No	No
48,733,102.00	No	No	48	Women	White	50.50	18.62	133.00	82.50	2.30	99.50	No	6.10	Yes	No	No
51,623,102.00	No	No	49	Women	White	71.00	18.63	168.50	97.50	1.90	121.50	No	4.80	No	No	No
71,718,102.00	No	No	57	Women	White	38.50	18.63	105.00	66.50	2.20	79.00	No	5.80	No	No	Yes
34,846,101.00	No	No	32	Women	White	44.50	18.65	118.50	74.00	1.20	89.00	No	7.00	No	No	Yes
20,023,102.00	No	No	77	Women	White	57.00	18.66	136.50	79.50	2.00	98.50	No	7.20	Yes	No	No

Table 5.9B: Absolute Percentage Risks and variable attributes values of Model I for the first 30 participants

Pserial no.	Grp	Bp1	Age	Sex	X1	X2	X3	X4	X5	X6	%PR
13,956,102.00	No	No	60	Women	50.0	25.5	0.02	0.00	50.00	5.22	-14.6
63,535,102.00	Yes	Yes	30	Women	45.8	25.9	4.15	0.00	100.19	4.28	-15.4
71,831,101.00	No	No	66	Women	0.0	25.5	140.00	0.00	100.00	0.00	-16.7
34,031,101.00	No	No	84	Women	50.0	25.5	0.00	0.00	100.00	5.30	-18.0
72,604,102.00	No	No	59	Women	50.0	25.5	0.00	0.00	100.00	0.00	-18.9
13,008,101.00	Yes	Yes	50	Women	50.0	25.5	0.00	0.00	100.00	0.00	-16.3
39,139,101.00	No	No	34	Women	50.0	25.5	0.00	0.00	100.00	0.00	-21.4
47,856,102.00	No	No	51	Women	50.0	25.5	0.01	0.00	100.00	5.23	-21.9
37,710,101.00	No	No	61	Women	50.0	25.5	0.00	0.00	100.00	0.00	-15.2
54,256,101.00	No	No	31	Women	50.0	25.5	0.00	0.00	100.00	5.30	-14.1
53,817,101.00	No	No	43	Women	50.0	25.5	0.00	0.00	100.00	5.30	-19.6
44,633,102.00	No	No	39	Women	50.0	25.4	0.04	0.00	100.01	2.67	-19.9
34,523,102.00	No	No	45	Women	0.0	25.5	0.00	0.00	100.00	5.30	-15.7
72,323,102.00	No	No	38	Women	50.0	25.5	0.00	0.00	100.00	5.30	-16.6
42,831,101.00	No	No	35	Men	0.0	25.5	0.00	0.00	100.00	5.20	-13.6
72,833,102.00	No	No	40	Men	0.0	25.5	0.00	0.00	100.00	5.20	-18.2
15,047,101.00	No	No	86	Women	0.0	25.5	0.00	0.00	100.00	0.00	-15.7
23,202,102.00	No	No	58	Women	50.0	25.5	0.00	0.00	100.00	5.30	-17.6
20,408,101.00	No	No	44	Women	50.0	25.5	0.00	0.00	100.00	5.30	-21.7
76,910,102.00	No	No	48	Women	50.0	25.5	0.00	0.00	100.00	0.00	-18.7
13,856,101.00	No	No	51	Women	50.0	25.5	0.00	0.00	100.00	0.00	-24.4
21,413,101.00	Yes	Yes	86	Men	10.3	21.9	111.29	0.00	34.49	2.95	-11.5
25,713,101.00	No	No	38	Women	50.0	25.5	0.00	0.00	100.00	5.30	-20.8
29,646,101.00	No	No	32	Women	50.0	25.5	0.00	0.00	100.00	5.30	-17.3
59,015,101.00	No	No	31	Women	50.0	25.5	0.00	0.00	100.00	5.30	-21.9
48,733,102.00	No	No	48	Women	0.0	25.5	0.00	0.00	100.00	0.00	-11.0
51,623,102.00	No	No	49	Women	0.0	25.5	140.00	0.00	0.00	5.30	-19.0
71,718,102.00	No	No	57	Women	50.0	25.5	0.00	0.00	100.00	0.00	-20.3
34,846,101.00	No	No	32	Women	50.0	25.5	0.00	0.00	100.00	0.00	-15.6
20,023,102.00	No	No	77	Women	0.0	25.5	0.00	0.00	100.00	0.00	-11.3

Table 5.10B: Calculation of TPR, FPR, LRP, and LRN, for the MATLAB Model I for the first 30 participants

Pserial no.	Grp	Bp1	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRM
13,956,102.00	No	No	60	Women	-14.6	0	1	1	0.9997	3574	0
63,535,102.00	Yes	Yes	30	Women	-15.4	0	1	1	0.9994	1787	0
71,831,101.00	No	No	66	Women	-16.7	0	1	1	0.9991	1191.33	0
34,031,101.00	No	No	84	Women	-18.0	0	1	1	0.9988	893.5	0
72,604,102.00	No	No	59	Women	-18.9	0	1	1	0.9985	714.8	0
13,008,101.00	Yes	Yes	50	Women	-16.3	0	1	1	0.9982	595.666	0
39,139,101.00	No	No	34	Women	-21.4	1	0	0.9966	0.9982	510.571	0
47,856,102.00	No	No	51	Women	-21.9	1	0	0.9932	0.9982	446.75	0
37,710,101.00	No	No	61	Women	-15.2	0	1	0.9932	0.9979	397.11	0
54,256,101.00	No	No	31	Women	-14.1	0	1	0.9932	0.9976	357.4	0
53,817,101.00	No	No	43	Women	-19.6	0	1	0.9932	0.9973	324.90	0
44,633,102.00	No	No	39	Women	-19.9	0	1	0.9932	0.9970	297.83	0
34,523,102.00	No	No	45	Women	-15.7	0	1	0.9932	0.9967	274.92	0
72,323,102.00	No	No	38	Women	-16.6	0	1	0.9932	0.9964	255.28	0
42,831,101.00	No	No	35	Men	-13.6	0	1	0.9932	0.9961	238.26	0
72,833,102.00	No	No	40	Men	-18.2	0	1	0.9932	0.9958	223.37	0
15,047,101.00	No	No	86	Women	-15.7	0	1	0.9932	0.9955	210.23	0
23,202,102.00	No	No	58	Women	-17.6	0	1	0.9932	0.9952	198.55	0
20,408,101.00	No	No	44	Women	-21.7	1	0	0.9898	0.9952	188.10	0
76,910,102.00	No	No	48	Women	-18.7	0	1	0.9898	0.9949	178.7	0
13,856,101.00	No	No	51	Women	-24.4	1	0	0.9865	0.9949	176.18	0.014
21,413,101.00	Yes	Yes	86	Men	-11.5	0	1	0.9865	0.9946	167.79	0.014
25,713,101.00	No	No	38	Women	-20.8	1	0	0.9831	0.9946	160.16	0.014
29,646,101.00	No	No	32	Women	-17.3	0	1	0.9831	0.9943	153.20	0.014
59,015,101.00	No	No	31	Women	-21.9	1	0	0.9797	0.9943	146.81	0.014
48,733,102.00	No	No	48	Women	-11.0	0	1	0.9797	0.9940	140.94	0.014
51,623,102.00	No	No	49	Women	-19.0	0	1	0.9797	0.9937	135.52	0.014
71,718,102.00	No	No	57	Women	-20.3	1	0	0.9764	0.9937	130.50	0.014
34,846,101.00	No	No	32	Women	-15.6	0	1	0.9764	0.9934	125.84	0.014
20,023,102.00	No	No	77	Women	-11.3	0	1	0.9764	0.9931	121.50	0.014

Table 5.11B: Initial Absolute Percentage Risks and attributes variable values for the first 30 participants Model II

Pserial no.	Grp	Bp1	Age	Sex	X1	X2	X3	X4	X5	X6	%PR
13,956,102.00	No	No	60	Women	50.1	22.6	31.4	1.0	50.3	3.3	12.8
63,535,102.00	Yes	Yes	30	Women	45.1	26.3	4.0	1.0	99.5	4.1	14.4
71,831,101.00	No	No	66	Women	10.3	14.1	102.7	1.0	19.2	6.5	12.2
34,031,101.00	No	No	84	Women	30.8	19.4	30.2	1.0	88.8	3.4	14.4
72,604,102.00	No	No	59	Women	35.7	9.1	34.4	1.0	76.7	6.5	13.4
13,008,101.00	Yes	Yes	50	Women	41.6	21.7	25.4	1.0	89.3	2.9	13.5
39,139,101.00	No	No	34	Women	36.1	21.3	2.4	1.0	100.2	2.7	19.1
47,856,102.00	No	No	51	Women	41.9	21.4	3.8	1.0	100.1	3.4	19.4
37,710,101.00	No	No	61	Women	48.9	24.4	3.4	1.0	98.7	5.4	14.1
54,256,101.00	No	No	31	Women	42.0	19.7	33.6	1.0	68.1	3.9	11.7
53,817,101.00	No	No	43	Women	45.0	20.3	22.6	1.0	99.3	3.5	16.3
44,633,102.00	No	No	39	Women	48.8	22.0	1.7	1.0	100.3	3.2	17.9
34,523,102.00	No	No	45	Women	19.7	19.4	31.3	1.0	87.0	3.5	12.9
72,323,102.00	No	No	38	Women	50.6	23.7	3.2	1.0	99.1	4.3	15.2
42,831,101.00	No	No	35	Men	10.4	20.8	41.0	1.0	82.3	3.9	11.6
72,833,102.00	No	No	40	Men	6.5	21.6	1.9	1.0	100.2	3.6	16.5
15,047,101.00	No	No	86	Women	11.2	17.8	53.6	1.0	78.6	2.5	12.2
23,202,102.00	No	No	58	Women	37.4	22.1	1.4	1.0	100.2	4.4	16.0
20,408,101.00	No	No	44	Women	48.1	21.0	1.7	1.0	100.3	3.5	19.4
76,910,102.00	No	No	48	Women	35.0	19.0	27.5	1.0	93.7	2.5	15.0
13,856,101.00	No	No	51	Women	8.1	10.6	115.0	1.0	0.0	3.3	1.8
21,413,101.00	Yes	Yes	86	Men	15.4	8.8	91.9	1.0	46.6	6.4	10.1
25,713,101.00	No	No	38	Women	50.6	20.9	1.9	1.0	99.9	4.0	18.7
29,646,101.00	No	No	32	Women	50.6	22.4	2.0	1.0	99.2	3.8	15.9
59,015,101.00	No	No	31	Women	48.2	20.8	1.7	1.0	100.3	3.7	19.6
48,733,102.00	No	No	48	Women	23.2	18.5	41.3	1.0	56.9	3.0	10.0
51,623,102.00	No	No	49	Women	10.8	18.8	123.2	1.0	19.9	3.5	14.2
71,718,102.00	No	No	57	Women	40.6	18.3	27.9	1.0	89.0	3.1	15.7
34,846,101.00	No	No	32	Women	47.8	23.6	2.1	1.0	100.2	2.3	14.4
20,023,102.00	No	No	77	Women	15.3	19.3	45.8	1.0	63.5	2.6	10.0

Table 5.12B: Calculation of TPR, FPR, LRP, and LRN, for the MATLAB Model II for the first 30 participants

Pserial no.	Grp	Bp1	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRM
13,956,102.00	No	No	60	Women	12.8	0	1	1	1.000	3571.43	0
63,535,102.00	Yes	Yes	30	Women	14.4	0	1	1	0.999	1785.71	0
71,831,101.00	No	No	66	Women	12.2	0	1	1	0.999	1191.90	0
34,031,101.00	No	No	84	Women	14.4	0	1	1	0.999	893.66	0
72,604,102.00	No	No	59	Women	13.4	0	1	1	0.999	714.80	0
13,008,101.00	Yes	Yes	50	Women	13.5	0	1	1	0.998	595.59	0
39,139,101.00	No	No	34	Women	19.1	0	1	1	0.998	510.46	0
47,856,102.00	No	No	51	Women	19.4	0	1	1	0.998	446.83	0
37,710,101.00	No	No	61	Women	14.1	0	1	1	0.997	397.14	0
54,256,101.00	No	No	31	Women	11.7	0	1	1	0.997	357.40	0
53,817,101.00	No	No	43	Women	16.3	0	1	1	0.997	324.89	0
44,633,102.00	No	No	39	Women	17.9	0	1	1	0.997	297.80	0
34,523,102.00	No	No	45	Women	12.9	0	1	1	0.996	274.95	0
72,323,102.00	No	No	38	Women	15.2	0	1	1	0.996	255.30	0
42,831,101.00	No	No	35	Men	11.6	0	1	1	0.996	238.27	0
72,833,102.00	No	No	40	Men	16.5	0	1	1	0.996	223.36	0
15,047,101.00	No	No	86	Women	12.2	0	1	1	0.995	210.22	0
23,202,102.00	No	No	58	Women	16.0	0	1	1	0.995	198.57	0
20,408,101.00	No	No	44	Women	19.4	0	1	1	0.995	188.11	0
76,910,102.00	No	No	48	Women	15.0	0	1	1	0.994	178.70	0
13,856,101.00	No	No	51	Women	1.8	1	0	0.986	0.994	176.18	0.0142
21,413,101.00	Yes	Yes	86	Men	10.1	0	1	0.986	0.994	167.79	0.0142
25,713,101.00	No	No	38	Women	18.7	0	1	0.986	0.994	160.16	0.0142
29,646,101.00	No	No	32	Women	15.9	0	1	0.986	0.994	153.21	0.0142
59,015,101.00	No	No	31	Women	19.6	0	1	0.986	0.993	146.82	0.0142
48,733,102.00	No	No	48	Women	10.0	0	1	0.986	0.993	140.95	0.0142
51,623,102.00	No	No	49	Women	14.2	0	1	0.986	0.993	135.52	0.0142
71,718,102.00	No	No	57	Women	15.7	0	1	0.986	0.992	130.50	0.0142
34,846,101.00	No	No	32	Women	14.4	0	1	0.986	0.992	125.85	0.0142
20,023,102.00	No	No	77	Women	10.0	0	1	0.986	0.992	121.51	0.0142

Appendix 6.0 for Chapter 6:

Appendix 6.1

CMAUT Complementary system with three organs

Example 2: Similar procedure was used to convert the CMAUT expression with three organs that is $[(Q_1 \text{ AND } Q_2 \text{ AND } Q_3), P_1, V_1, P_2, V_2, P_3, V_3]$ into a set of inequalities. First convert the attributes $P_1, V_1, P_2, V_2, P_3, V_3$ into common attribute utility units (U_1 and U_2 and U_3) using the utility formula. $[(Q_1 \text{ AND } Q_2 \text{ AND } Q_3) \text{ equiv } U_4]$ into CNF and then to a set of inequalities (or constraints). The results of simplifying the clauses using the above algorithm are as follows:

$$((Q_1 \text{ OR NOT } U_4) \text{ AND } (Q_2 \text{ OR NOT } U_4) \text{ AND } (Q_3 \text{ OR NOT } U_4) \text{ AND } (U_4 \text{ OR NOT } Q_1 \text{ OR NOT } Q_2 \text{ OR NOT } Q_3)) \quad (6.6)$$

The numbers of Clauses that are obtained from the above CMAUT expression (6.6) is:

$$((Q_1 \text{ OR NOT } U_4) \text{ AND } (Q_2 \text{ OR NOT } U_4) \text{ AND } (Q_3 \text{ OR NOT } U_4) \text{ AND } (U_4 \text{ OR NOT } Q_1 \text{ OR NOT } Q_2 \text{ OR NOT } Q_3))$$

To convert the first clause in the expression (6.6) into a set of inequalities, the following operation is carried out.

Again, the results of transforming the clauses in the CNF above into a set of inequalities using Table 6.1 are:

Consider the first three clauses in the expression (6.6) and **convert them into** set of inequalities;

$$\begin{aligned} (Q_1 \text{ OR NOT } U_4) &= Q_1 \vee \neg U_4; -q_1 + u_4 \leq 0; q_1 - u_4 \geq 0 \\ (Q_2 \text{ OR NOT } U_4) &= Q_2 \vee \neg U_4; -q_2 + u_4 \leq 0; q_2 - u_4 \geq 0 \\ (Q_3 \text{ OR NOT } U_4) &= Q_3 \vee \neg U_4; -q_3 + u_4 \leq 0; q_3 - u_4 \geq 0; \end{aligned}$$

Consider the fourth clause in the expression (6.6) and **convert it into** a set of inequalities

$$\begin{aligned} (U_4 \text{ OR NOT } Q_1 \text{ OR NOT } Q_2 \text{ OR NOT } Q_3) \\ (U_4 \vee \neg Q_1 \vee \neg Q_2 \vee \neg Q_3) &\geq 1 \\ (1 + u_4 - q_1) \vee \neg Q_2 \vee \neg Q_3 &\geq 1 \\ (1 + 1 + u_4 - q_1 - q_2) \vee \neg Q_3 &\geq 1 \\ 3 + u_4 - q_1 - q_2 - q_3 &\geq 1 \\ q_1 + q_2 + q_3 - u_4 &\leq 2 \end{aligned}$$

Putting all the set of inequalities from the example 2 and expression (6.6) together, give the following:

$$\begin{aligned} q_1 - u_1 &\geq 0 \\ q_2 - u_1 &\geq 0 \\ q_3 - u_1 &\geq 0 \\ q_1 + q_2 + q_3 - u_1 &\leq 2 \end{aligned}$$

Other examples of how the above algorithm and Table 6.1 were used to solve and convert the CMAUT expressions into set of inequalities are shown in appendix 6.

Non- CMAUT Complementary system with three organs

Example 4: For the purpose of comparison the blood pressure is set to be equivalent of the utility unit of the organ. This means the utility unit U_i can take either blood pressure or blood volume values. In this example the expression used is: - $[(Q_1 \text{ equiv } P_1) \text{ AND } (Q_2 \text{ equiv } P_2) \text{ AND } (Q_3 \text{ equiv } P_3)]$ is used and also instead of P_i the U_i is used and result of eliminating high level operations is:

$$(((Q_1 \text{ AND } U_1) \text{ OR } (\text{NOT } Q_1 \text{ AND } \text{NOT } U_1)) \text{ AND } ((Q_2 \text{ AND } U_2) \text{ OR } (\text{NOT } Q_2 \text{ AND } \text{NOT } U_2))) \text{ AND } ((Q_3 \text{ AND } U_3) \text{ OR } (\text{NOT } Q_3 \text{ AND } \text{NOT } U_3)))$$

The next stage using the algorithm is pushing the NOT connector to the innermost level and this gives the following result:

$$(((Q_1 \text{ AND } U_1) \text{ OR } (\text{NOT } Q_1 \text{ AND } \text{NOT } U_1)) \text{ AND } ((Q_2 \text{ AND } U_2) \text{ OR } (\text{NOT } Q_2 \text{ AND } \text{NOT } U_2))) \text{ AND } ((Q_3 \text{ AND } U_3) \text{ OR } (\text{NOT } Q_3 \text{ AND } \text{NOT } U_3)))$$

The next step is to push OR operator to the innermost level which gives the results:

$$((Q_1 \text{ OR } \text{NOT } Q_1) \text{ AND } (Q_1 \text{ OR } \text{NOT } U_1) \text{ AND } (U_1 \text{ OR } \text{NOT } Q_1) \text{ AND } (U_1 \text{ OR } \text{NOT } U_1) \text{ AND } (Q_2 \text{ OR } \text{NOT } Q_2) \text{ AND } (Q_2 \text{ OR } \text{NOT } U_2) \text{ AND } (U_2 \text{ OR } \text{NOT } Q_2) \text{ AND } (U_2 \text{ OR } \text{NOT } U_2) \text{ AND } (Q_3 \text{ OR } \text{NOT } Q_3) \text{ AND } (Q_3 \text{ OR } \text{NOT } U_3) \text{ AND } (U_3 \text{ OR } \text{NOT } Q_3) \text{ AND } (U_3 \text{ OR } \text{NOT } U_3)).$$

The final list of clauses after simplification is as follows:

$$((Q_1 \text{ OR } \text{NOT } U_1) \text{ AND } (U_1 \text{ OR } \text{NOT } Q_1) \text{ AND } (Q_2 \text{ OR } \text{NOT } U_2) \text{ AND } (U_2 \text{ OR } \text{NOT } Q_2) \text{ AND } (Q_3 \text{ OR } \text{NOT } U_3) \text{ AND } (U_3 \text{ OR } \text{NOT } Q_3))$$

The number of clauses the expression $[(Q_1 \text{ equiv } U_1) \text{ AND } (Q_2 \text{ equiv } U_2) \text{ AND } (Q_3 \text{ equiv } U_3)]$ generates is as below:

1. $(Q_1 \text{ OR } \text{NOT } U_1)$
2. $(U_1 \text{ OR } \text{NOT } Q_1)$
3. $(Q_2 \text{ OR } \text{NOT } U_2)$
4. $(U_2 \text{ OR } \text{NOT } Q_2)$
5. $(Q_3 \text{ OR } \text{NOT } U_3)$
6. $(U_3 \text{ OR } \text{NOT } Q_3)$

To transform the clauses in the CNF into a set of inequalities, **Table 6.1** was used and the results are presented below as:

$$\begin{aligned} q_1 - u_1 &\geq 0 \\ u_1 - q_1 &\geq 0 \\ q_2 - u_2 &\geq 0 \\ u_2 - q_2 &\geq 0 \\ q_3 - u_3 &\geq 0 \\ u_3 - q_3 &\geq 0 \end{aligned}$$

It is note that in all these conversion the number of clauses in the CNF are equal to the number of inequalities the expression generates.

CMAUT substitutable system with three organs

Example 6: this second example converts three substitutable organs with a common utility unit into a set of inequalities. The example uses the expression $[(Q_1 \text{ OR } Q_2 \text{ OR } Q_3)P_1, V_1, P_2, V_2, P_3, V_3]$ to generate the constraints. This expression is rewritten in CMAUT format as

$$[(Q_1 \text{ equiv } P_1, V_1) \text{ OR } (Q_2 \text{ equiv } P_2, V_2) \text{ OR } (Q_3 \text{ equiv } P_3, V_3)]$$

The attributes in the expression are converted into utility units, which are U_1 , U_2 and U_3 . the finally expression is written as $[(Q_1 \text{ equiv } U_1) \text{ OR } (Q_2 \text{ equiv } U_2) \text{ OR } (Q_3 \text{ equiv } U_3)]$. Upon simplification the list of clauses will be as follows:

1. $(Q_1 \text{ OR } Q_2 \text{ OR } Q_3 \text{ OR NOT } U_1 \text{ OR NOT } U_2 \text{ OR NOT } U_3)$
2. $(Q_1 \text{ OR } Q_2 \text{ OR } U_3 \text{ OR NOT } U_1 \text{ OR NOT } G_3 \text{ OR NOT } U_2)$
3. $(Q_1 \text{ OR } Q_3 \text{ OR } U_2 \text{ OR NOT } Q_1 \text{ OR NOT } U_1 \text{ OR NOT } U_3)$
4. $(Q_1 \text{ OR } U_2 \text{ OR } U_3 \text{ OR NOT } Q_2 \text{ OR NOT } U_1 \text{ OR NOT } Q_3)$
5. $(Q_2 \text{ OR } U_1 \text{ OR } Q_3 \text{ OR NOT } Q_1 \text{ OR NOT } U_2 \text{ OR NOT } U_3)$
6. $(Q_2 \text{ OR } U_1 \text{ OR } U_3 \text{ OR NOT } Q_1 \text{ OR NOT } Q_3 \text{ OR NOT } U_2)$
7. $(U_1 \text{ OR } Q_3 \text{ OR } U_2 \text{ OR NOT } Q_1 \text{ OR NOT } Q_2 \text{ OR NOT } U_3)$
8. $(U_1 \text{ OR } U_2 \text{ OR } U_3 \text{ OR NOT } Q_1 \text{ OR NOT } Q_2 \text{ OR NOT } Q_3)$

The set of inequalities with integer variables would be:

$$\begin{aligned} q_1 + q_2 + q_3 - u_1 - u_2 - u_3 &\geq -2 \\ q_1 + q_2 + u_3 - u_1 - q_3 - u_2 &\geq -2 \\ q_1 + q_3 + u_2 - q_2 - u_1 - u_3 &\geq -2 \\ q_1 + u_1 + u_3 - q_2 - u_1 - q_3 &\geq -2 \\ q_2 + u_1 + q_3 - q_1 - u_2 - u_3 &\geq -2 \\ q_2 + u_1 + u_3 - q_1 - q_3 - u_2 &\geq -2 \\ u_1 + q_3 + u_2 - q_1 - q_2 - u_3 &\geq -2 \\ u_1 + u_2 + u_3 - q_1 - q_2 - q_3 &\geq -2 \end{aligned}$$

Non- CMAUT substitutable system with three organs (OR)

Example 8:- Similarly the *Non – CMAUT* with three organs can be expressed as $[(Q_1 \text{ equiv } P_1) \text{ OR } (Q_2 \text{ equiv } P_2) \text{ OR } (Q_3 \text{ equiv } P_3)]$ and the number of clauses obtained after simplification is as follows:

$$\begin{aligned} (Q_1 \text{ OR } Q_2 \text{ OR } Q_3 \text{ OR NOT } P_1 \text{ OR NOT } P_2 \text{ OR NOT } P_3) \\ (Q_1 \text{ OR } Q_2 \text{ OR } P_3 \text{ OR NOT } P_1 \text{ OR NOT } G_3 \text{ OR NOT } P_2) \\ (Q_1 \text{ OR } Q_3 \text{ OR } P_2 \text{ OR NOT } Q_1 \text{ OR NOT } P_1 \text{ OR NOT } P_3) \\ (Q_1 \text{ OR } P_2 \text{ OR } P_3 \text{ OR NOT } Q_2 \text{ OR NOT } P_1 \text{ OR NOT } Q_3) \\ (Q_2 \text{ OR } P_1 \text{ OR } Q_3 \text{ OR NOT } Q_1 \text{ OR NOT } P_2 \text{ OR NOT } P_3) \\ (Q_2 \text{ OR } P_1 \text{ OR } P_3 \text{ OR NOT } Q_1 \text{ OR NOT } Q_3 \text{ OR NOT } P_2) \\ (P_1 \text{ OR } Q_3 \text{ OR } P_2 \text{ OR NOT } Q_1 \text{ OR NOT } Q_2 \text{ OR NOT } P_3) \\ (P_1 \text{ OR } P_2 \text{ OR } P_3 \text{ OR NOT } Q_1 \text{ OR NOT } Q_2 \text{ OR NOT } Q_3) \end{aligned}$$

The set of inequalities with integer variables would be:

$$q_1 + q_2 + q_3 - p_1 - p_2 - p_3 \geq -2$$

$$q_1 + q_2 + p_3 - p_1 - q_3 - p_2 \geq -2$$

$$q_1 + q_3 + p_2 - q_2 - p_1 - p_3 \geq -2$$

$$q_1 + p_1 + p_3 - q_2 - p_1 - q_3 \geq -2$$

$$q_2 + p_1 + q_3 - q_1 - p_2 - p_3 \geq -2$$

$$q_2 + p_1 + p_3 - q_1 - q_3 - p_2 \geq -2$$

$$p_1 + q_3 + p_2 - q_1 - q_2 - p_3 \geq -2$$

$$p_1 + p_2 + p_3 - q_1 - q_2 - q_3 \geq -2$$

Comparing the above examples with those of the *CMAUT* CIS it can be inferred that the clauses and the set of inequalities are the same. Thus for substitutable organs, *CMAUT*, where OR is used, the same expression can be used for both *CMAUT* and *Non – CMAUT* CIS.

- In Appendix 9 is Table 9.1 for the comparison of the inequalities constraints generated for substitutable and complementary data representation using *CMAUT* and *Non- CMAUT*:
- Table 6.6C: Data sizes of the 402 participants before and after optimisation with *CMAUT* framework in electronic format:

Table 6.6B: Data sizes for first 30 participants before and after optimisation with *CMAUT*

No. Of participants	Pserial no.	Data size before optimisation (bytes)	Data size after optimisation (bytes)
1	10,902,101.00	1256	465
2	10,846,103.00	1251	464
3	11,039,102.00	1251	463
4	11,046,101.00	1251	465
5	11,239,101.00	1245	464
6	11,249,102.00	1244	464
7	11,306,101.00	1249	464
8	11,313,101.00	1245	463
9	11,349,102.00	1243	464
10	11,356,101.00	1262	464
11	11,410,102.00	1260	463
12	11,410,101.00	1241	465
13	11,435,101.00	1253	463
14	11,439,101.00	1231	464
15	11,449,101.00	1244	463
16	11,449,102.00	1266	464
17	11,506,102.00	1606	463
18	11,547,101.00	1222	464
19	11,610,102.00	1256	463
20	11,610,101.00	1248	462
21	11,618,101.00	1273	463
22	11,633,101.00	1269	465
23	11,649,101.00	1230	463
24	11,714,102.00	1245	465
25	10,904,101.00	1234	465
26	10,904,102.00	1252	463
27	10,908,101.00	1242	464
28	10,913,101.00	1229	464
29	10,915,101.00	1266	463
30	10,927,101.00	1268	463

Appendix 7.0 for Chapter 7:

Table 7.3B: The Raw data of the first 30 participants used in Chapter 7:

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
71,831,101.00	No	No	66	Women	White	89.00	14.32	159.00	70.00	1.90	99.50	No	6.90	No	No	Yes
34,031,101.00	No	No	84	Women	White	48.50	16.17	112.00	63.50	2.20	80.00	No	5.00	Yes	Yes	Yes
72,604,102.00	No	No	59	Women	White	36.50	16.19	109.50	73.00	2.00	85.00	No	6.00	No	No	No
13,008,101.00	Yes	Yes	50	Women	White	43.00	16.65	117.00	74.00	1.70	88.50	No	6.00	Yes	Yes	Yes
39,139,101.00	No	No	34	Women	White	48.00	16.81	102.00	54.00	1.80	70.00	No	6.50	Yes	No	No
47,856,102.00	No	No	51	Women	White	44.00	16.85	100.50	56.50	1.90	71.00	No	5.10	No	No	No
37,710,101.00	No	No	61	Women	White	43.00	17.43	120.00	77.00	1.20	91.50	No	5.50	No	No	No
54,256,101.00	No	No	31	Women	White	40.00	17.72	124.00	84.00	2.00	97.00	No	3.90	Yes	No	No
53,817,101.00	No	No	43	Women	White	39.00	17.72	107.00	68.00	1.70	81.00	No	4.90	Yes	No	No
44,633,102.00	No	No	39	Women	White	39.00	17.87	106.00	67.00	1.40	80.00	No	5.30	No	No	No
34,523,102.00	No	No	45	Women	White	51.50	18.09	120.00	68.50	2.00	85.50	No	4.70	No	No	No
72,323,102.00	No	No	38	Women	White	39.00	18.10	116.00	77.00	1.40	90.00	No	4.20	Yes	Yes	No
42,831,101.00	No	No	35	Men	White	64.50	18.15	135.00	70.50	1.70	92.00	No	4.10	Yes	No	No
72,833,102.00	No	No	40	Men	White	58.50	18.19	117.00	58.50	1.50	78.00	No	4.60	Yes	No	No
15,047,101.00	No	No	86	Women	White	79.00	18.27	138.00	59.00	2.50	85.50	No	8.90	Yes	No	Yes
23,202,102.00	No	No	58	Women	White	48.00	18.29	113.50	65.50	1.10	81.50	No	3.60	Yes	No	No
20,408,101.00	No	No	44	Women	White	38.00	18.41	100.00	62.00	0.90	75.00	No	4.90	No	No	No
76,910,102.00	No	No	48	Women	White	46.00	18.43	110.00	64.00	2.00	79.00	No	8.00	Yes	No	No
13,856,101.00	No	No	51	Women	White	32.50	18.45	92.50	60.00	2.60	70.50	No	6.70	Yes	No	Yes
21,413,101.00	Yes	Yes	86	Men	White	65.50	18.48	149.00	83.50	1.20	105.00	No	5.60	No	Yes	No
25,713,101.00	No	No	38	Women	White	32.50	18.55	103.50	71.00	1.30	81.50	No	4.00	Yes	No	No
29,646,101.00	No	No	32	Women	White	36.00	18.58	113.50	77.50	1.40	89.50	No	4.60	No	No	No
59,015,101.00	No	No	31	Women	Asian or Asian British	38.00	18.59	100.00	62.00	1.50	74.50	No	4.50	Yes	No	No
48,733,102.00	No	No	48	Women	White	50.50	18.62	133.00	82.50	2.30	99.50	No	6.10	Yes	No	No
51,623,102.00	No	No	49	Women	White	71.00	18.63	168.50	97.50	1.90	121.50	No	4.80	No	No	No
71,718,102.00	No	No	57	Women	White	38.50	18.63	105.00	66.50	2.20	79.00	No	5.80	No	No	Yes
34,846,101.00	No	No	32	Women	White	44.50	18.65	118.50	74.00	1.20	89.00	No	7.00	No	No	Yes
20,023,102.00	No	No	77	Women	White	57.00	18.66	136.50	79.50	2.00	98.50	No	7.20	Yes	No	No

Table 7.4B: Predicative Percentage Risks for 10 years and attribute variable values for the first 30 participants (from Model I 3645 data sets)

Pserial no.	Grp	Bp1	Age	Sex	X1	X2	X3	X4	X5	X6	%PR
13,956,102.00	No	No	60	Women	50.0	25.5	0.02	0.00	50.00	5.22	14.9
63,535,102.00	Yes	Yes	30	Women	45.8	25.9	4.15	0.00	100.19	4.28	16.3
71,831,101.00	No	No	66	Women	0.0	25.5	140.00	0.00	100.00	0.00	17.1
34,031,101.00	No	No	84	Women	50.0	25.5	0.00	0.00	100.00	5.30	19.0
72,604,102.00	No	No	59	Women	50.0	25.5	0.00	0.00	100.00	0.00	19.2
13,008,101.00	Yes	Yes	50	Women	50.0	25.5	0.00	0.00	100.00	0.00	17.3
39,139,101.00	No	No	34	Women	50.0	25.5	0.00	0.00	100.00	0.00	21.7
47,856,102.00	No	No	51	Women	50.0	25.5	0.01	0.00	100.00	5.23	22.2
37,710,101.00	No	No	61	Women	50.0	25.5	0.00	0.00	100.00	0.00	15.6
54,256,101.00	No	No	31	Women	50.0	25.5	0.00	0.00	100.00	5.30	14.4
53,817,101.00	No	No	43	Women	50.0	25.5	0.00	0.00	100.00	5.30	19.9
44,633,102.00	No	No	39	Women	50.0	25.4	0.04	0.00	100.01	2.67	20.2
34,523,102.00	No	No	45	Women	0.0	25.5	0.00	0.00	100.00	5.30	16.1
72,323,102.00	No	No	38	Women	50.0	25.5	0.00	0.00	100.00	5.30	17.5
42,831,101.00	No	No	35	Men	0.0	25.5	0.00	0.00	100.00	5.20	13.9
72,833,102.00	No	No	40	Men	0.0	25.5	0.00	0.00	100.00	5.20	18.5
15,047,101.00	No	No	86	Women	0.0	25.5	0.00	0.00	100.00	0.00	16.0
23,202,102.00	No	No	58	Women	50.0	25.5	0.00	0.00	100.00	5.30	17.9
20,408,101.00	No	No	44	Women	50.0	25.5	0.00	0.00	100.00	5.30	22.0
76,910,102.00	No	No	48	Women	50.0	25.5	0.00	0.00	100.00	0.00	19.0
13,856,101.00	No	No	51	Women	50.0	25.5	0.00	0.00	100.00	0.00	24.6
21,413,101.00	Yes	Yes	86	Men	10.3	21.9	111.29	0.00	34.49	2.95	11.7
25,713,101.00	No	No	38	Women	50.0	25.5	0.00	0.00	100.00	5.30	21.1
29,646,101.00	No	No	32	Women	50.0	25.5	0.00	0.00	100.00	5.30	17.6
59,015,101.00	No	No	31	Women	50.0	25.5	0.00	0.00	100.00	5.30	22.2
48,733,102.00	No	No	48	Women	0.0	25.5	0.00	0.00	100.00	0.00	11.4
51,623,102.00	No	No	49	Women	0.0	25.5	140.00	0.00	0.00	5.30	19.5
71,718,102.00	No	No	57	Women	50.0	25.5	0.00	0.00	100.00	0.00	20.5
34,846,101.00	No	No	32	Women	50.0	25.5	0.00	0.00	100.00	0.00	15.9
20,023,102.00	No	No	77	Women	0.0	25.5	0.00	0.00	100.00	0.00	14.9

Table 7.5B: Calculation of TPR, FPR, LRP, and LRN, for the MATLAB Model I PPR for 10 years for the first 30 participants (from Model I 3645 data sets)

Pserial no.	Grp	Bpl	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRN
13,956,102.00	No	No	60	Women	15.6	0	1	1	1.000	3278.7	0
63,535,102.00	Yes	Yes	30	Women	16.3	0	1	1	0.999	1639.3	0
71,831,101.00	No	No	66	Women	17.7	0	1	1	0.999	1092.9	0
34,031,101.00	No	No	84	Women	19.0	0	1	1	0.999	819.7	0
72,604,102.00	No	No	59	Women	19.8	0	1	1	0.998	655.7	0
13,008,101.00	Yes	Yes	50	Women	21.0	0	1	1	0.998	546.4	0
39,139,101.00	No	No	34	Women	19.4	1	0	0.997	0.998	545.0	0.0027
47,856,102.00	No	No	51	Women	17.2	1	0	0.995	0.998	543.5	0.0055
37,710,101.00	No	No	61	Women	18.3	0	1	0.995	0.998	465.8	0.0055
54,256,101.00	No	No	31	Women	22.2	0	1	0.995	0.998	407.4	0.0055
53,817,101.00	No	No	43	Women	22.8	0	1	0.995	0.997	362.2	0.0055
44,633,102.00	No	No	39	Women	20.2	1	0	0.992	0.997	361.2	0.0082
34,523,102.00	No	No	45	Women	19.2	0	1	0.992	0.997	325.1	0.0082
72,323,102.00	No	No	38	Women	20.4	0	1	0.992	0.997	295.5	0.0082
42,831,101.00	No	No	35	Men	16.2	0	1	0.992	0.996	270.9	0.0082
72,833,102.00	No	No	40	Men	21.5	0	1	0.992	0.996	250.1	0.0082
15,047,101.00	No	No	86	Women	20.0	0	1	0.992	0.996	232.2	0.0082
23,202,102.00	No	No	58	Women	18.2	0	1	0.992	0.995	216.7	0.0082
20,408,101.00	No	No	44	Women	18.0	1	0	0.989	0.995	216.1	0.0109
76,910,102.00	No	No	48	Women	18.1	0	1	0.989	0.995	202.6	0.0110
13,856,101.00	No	No	51	Women	18.4	1	0	0.986	0.995	202.1	0.0137
21,413,101.00	Yes	Yes	86	Men	15.0	0	1	0.986	0.995	190.2	0.0137
25,713,101.00	No	No	38	Women	20.5	1	0	0.984	0.995	189.7	0.0164
29,646,101.00	No	No	32	Women	15.5	0	1	0.984	0.995	179.1	0.0164
59,015,101.00	No	No	31	Women	22.0	1	0	0.981	0.995	178.6	0.0192
48,733,102.00	No	No	48	Women	20.8	0	1	0.981	0.994	169.2	0.0192
51,623,102.00	No	No	49	Women	23.8	0	1	0.981	0.994	160.8	0.0192
71,718,102.00	No	No	57	Women	21.9	1	0	0.978	0.994	160.3	0.0219
34,846,101.00	No	No	32	Women	16.7	0	1	0.978	0.994	152.7	0.0219
20,023,102.00	No	No	77	Women	11.8	0	1	0.978	0.993	145.8	0.0219

Table 7.13B: Predicative Percentage Risks for 10 years and attribute variable values for the first 30 participants Model II

Pserial no.	Grp	Bp1	Age	Sex	X1	X2	X3	X4	X5	X6	%PPR
13,956,102.00	No	No	60	Women	50.0	25.5	0.02	0.00	50.00	5.22	14.79
63,535,102.00	Yes	Yes	30	Women	45.8	25.9	4.15	0.00	100.19	4.28	15.53
71,831,101.00	No	No	66	Women	0.0	25.5	140.00	0.00	100.00	0.00	16.42
34,031,101.00	No	No	84	Women	50.0	25.5	0.00	0.00	100.00	5.30	17.64
72,604,102.00	No	No	59	Women	50.0	25.5	0.00	0.00	100.00	0.00	18.26
13,008,101.00	Yes	Yes	50	Women	50.0	25.5	0.00	0.00	100.00	0.00	16.13
39,139,101.00	No	No	34	Women	50.0	25.5	0.00	0.00	100.00	0.00	20.28
47,856,102.00	No	No	51	Women	50.0	25.5	0.01	0.00	100.00	5.23	20.75
37,710,101.00	No	No	61	Women	50.0	25.5	0.00	0.00	100.00	0.00	15.26
54,256,101.00	No	No	31	Women	50.0	25.5	0.00	0.00	100.00	5.30	14.26
53,817,101.00	No	No	43	Women	50.0	25.5	0.00	0.00	100.00	5.30	18.81
44,633,102.00	No	No	39	Women	50.0	25.4	0.04	0.00	100.01	2.67	19.02
34,523,102.00	No	No	45	Women	0.0	25.5	0.00	0.00	100.00	5.30	15.65
72,323,102.00	No	No	38	Women	50.0	25.5	0.00	0.00	100.00	5.30	16.33
42,831,101.00	No	No	35	Men	0.0	25.5	0.00	0.00	100.00	5.20	13.86
72,833,102.00	No	No	40	Men	0.0	25.5	0.00	0.00	100.00	5.20	17.68
15,047,101.00	No	No	86	Women	0.0	25.5	0.00	0.00	100.00	0.00	15.67
23,202,102.00	No	No	58	Women	50.0	25.5	0.00	0.00	100.00	5.30	17.18
20,408,101.00	No	No	44	Women	50.0	25.5	0.00	0.00	100.00	5.30	20.58
76,910,102.00	No	No	48	Women	50.0	25.5	0.00	0.00	100.00	0.00	18.08
13,856,101.00	No	No	51	Women	50.0	25.5	0.00	0.00	100.00	0.00	22.76
21,413,101.00	Yes	Yes	86	Men	10.3	21.9	111.29	0.00	34.49	2.95	13.61
25,713,101.00	No	No	38	Women	50.0	25.5	0.00	0.00	100.00	5.30	19.57
29,646,101.00	No	No	32	Women	50.0	25.5	0.00	0.00	100.00	5.30	16.91
59,015,101.00	No	No	31	Women	50.0	25.5	0.00	0.00	100.00	5.30	20.68
48,733,102.00	No	No	48	Women	0.0	25.5	0.00	0.00	100.00	0.00	11.73
51,623,102.00	No	No	49	Women	0.0	25.5	140.00	0.00	0.00	5.30	18.14
71,718,102.00	No	No	57	Women	50.0	25.5	0.00	0.00	100.00	0.00	19.39
34,846,101.00	No	No	32	Women	50.0	25.5	0.00	0.00	100.00	0.00	15.54
20,023,102.00	No	No	77	Women	0.0	25.5	0.00	0.00	100.00	0.00	12.01

Table 7.14B: Calculation of TPR, FPR, LRP, and LRN, for the MATLAB Model II for 10 years for the PPR first 30 participants

Pserial no.	Grp	Bpl	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRM
13,956,102.00	No	No	60	Women	14.79	0	1	1	0.9997	3484.32	0
63,535,102.00	Yes	Yes	30	Women	15.53	0	1	1	0.9994	1745.20	0
71,831,101.00	No	No	66	Women	16.42	0	1	1	0.9991	1162.79	0
34,031,101.00	No	No	84	Women	17.64	0	1	1	0.9989	872.60	0
72,604,102.00	No	No	59	Women	18.26	0	1	1	0.9986	697.84	0
13,008,101.00	Yes	Yes	50	Women	16.13	0	1	1	0.9983	581.73	0
39,139,101.00	No	No	34	Women	20.28	1	0	0.9935	0.9983	577.98	0.0065
47,856,102.00	No	No	51	Women	20.75	1	0	0.9871	0.9983	574.23	0.0129
37,710,101.00	No	No	61	Women	15.26	0	1	0.9871	0.9980	492.07	0.0129
54,256,101.00	No	No	31	Women	14.26	0	1	0.9871	0.9977	430.67	0.0129
53,817,101.00	No	No	43	Women	18.81	0	1	0.9871	0.9974	382.74	0.0129
44,633,102.00	No	No	39	Women	19.02	0	1	0.9871	0.9971	344.54	0.0129
34,523,102.00	No	No	45	Women	15.65	0	1	0.9871	0.9968	313.17	0.0129
72,323,102.00	No	No	38	Women	16.33	0	1	0.9871	0.9966	287.11	0.0129
42,831,101.00	No	No	35	Men	13.86	0	1	0.9871	0.9963	264.99	0.0130
72,833,102.00	No	No	40	Men	17.68	0	1	0.9871	0.9960	246.10	0.0130
15,047,101.00	No	No	86	Women	15.67	0	1	0.9871	0.9957	229.66	0.0130
23,202,102.00	No	No	58	Women	17.18	0	1	0.9871	0.9954	215.29	0.0130
20,408,101.00	No	No	44	Women	20.58	1	0	0.9806	0.9954	213.88	0.0194
76,910,102.00	No	No	48	Women	18.08	0	1	0.9806	0.9951	201.32	0.0194
13,856,101.00	No	No	51	Women	22.76	1	0	0.9742	0.9951	200.00	0.0259
21,413,101.00	Yes	Yes	86	Men	13.61	0	1	0.9742	0.9948	188.87	0.0259
25,713,101.00	No	No	38	Women	19.57	0	1	0.9742	0.9946	178.95	0.0259
29,646,101.00	No	No	32	Women	16.91	0	1	0.9742	0.9943	169.99	0.0260
59,015,101.00	No	No	31	Women	20.68	1	0	0.9677	0.9943	168.86	0.0324
48,733,102.00	No	No	48	Women	11.73	0	1	0.9677	0.9940	160.83	0.0325
51,623,102.00	No	No	49	Women	18.14	0	1	0.9677	0.9937	153.51	0.0325
71,718,102.00	No	No	57	Women	19.39	0	1	0.9677	0.9934	146.85	0.0325
34,846,101.00	No	No	32	Women	15.54	0	1	0.9677	0.9931	140.72	0.0325
20,023,102.00	No	No	77	Women	12.01	0	1	0.9677	0.9928	135.10	0.0325

Appendix 8.0 for Chapter 8

Appendix 8.1

1. [BHS - Cardiovascular Risk Charts and Calculators](#)

Cardiovascular Risk Charts and Calculators. The JBS CV Risk Assessor software is available to download from the Heart UK website with kind permission of ...

www.bhsoc.org/Cardiovascular_Risk_Charts_and_Calculators.stm - [Cached](#)

2. [Primary Cardiovascular Risk Calculator | Doctor | Patient UK](#)

13 Jan 2011 – Primary Cardiovascular Risk Calculator - Cardiovascular Risk Calculator For Primary Prevention This calculator should not be used if patient ...

www.patient.co.uk › [PatientPlus](#) - [Cached](#) - [Similar](#)

3. [Cardiovascular Risk Calculator and Chart v3.0](#)

28 May 2010 – There are also a number of alternative graphical displays that may be useful in discussing cardiovascular risk. There are a number of ...

4. [Calculator](#) - [Guidelines](#) - [Contacts](#) cvrisk.mvm.ed.ac.uk/ - [Cached](#) - [Similar](#)

5. [Cardiac Risk Calculator](#)

The standalone Cardiac Risk Calculator from. Age Gender F M BP / T-chol. HDL ...

www.cvhealth.ed.ac.uk/othercalcs/cardiacrisk.html - [Show more results from ed.ac.uk](#)

6. [QRISK2-2011](#)

Welcome to the QRISK®2-2011 risk calculator: <http://qrisk.org>. This calculator is only valid if you do not already have a diagnosis. ... Welcome to the QRISK®2-2011 cardiovascular disease risk calculator. Welcome to the QRISK®2-2011 Web ...

qrisk.org/ - [Cached](#)

7. [Cardiac risk calculators](#)

Equates to a risk of complications (ie cardiac death, documented intra- or postoperative MI, pulmonary oedema or nonfatal ventricular tachycardia) ...

www.vasgbi.com/riskdetsky.htm - [Cached](#) - [Similar](#)

8. [10-year CVD Risk Calculator \(Risk Assessment Tool for Estimating ...](#)

This tool is designed to estimate risk in adults aged 20 and older who do not have heart disease or diabetes. Use the calculator below to estimate 10-year ...

hp2010.nhlbihin.net/atpiii/calculator.asp?usertype=prof - [Cached](#) - [Similar](#)

9. [Cardiovascular risk score](#)

13 Jun 2006 – A Risk Score for Cardiovascular Disease ... For any technical issues with the calculator or these web pages, please e-mail ...

www.riskscore.org.uk/ - [Cached](#) - [Similar](#)

10. [CVD risk calculators - Clinical Knowledge Summaries](#)

Note that the QRISK® cardiovascular lifetime risk calculator is different and does not calculate 10-year risk of cardiovascular disease. ...

www.cks.nhs.uk/cvd_risk...risk/cvd_risk_calculators - [Cached](#) - [Similar](#)

Appendix 8.2:Output screen shots from the old Version 24 (2009) of the www.patient.co.uk website:

Cardiovascular Risk Calculator For Primary Prevention

This calculator should not be used if patient has known CVD or diabetes (already known to be at high risk)

Age (30-74)	30	Smoking Status	Smoker
Sex	Female	Glucose	Normal
Systolic BP	120	LVH	No LVH
Diastolic BP	74	Central Obesity	No
Total Cholesterol	4.5	South Asian Origin	No
HDL Cholesterol	1.4	Family History of CVD (Men <55 and women <65 years)	Significant FH CVD
Total /HDL Ratio	3.21	<input type="button" value="Calculate"/> <input type="button" value="Clear Fields"/>	
Serum TG mmol/L			

Using Systolic BP prediction, the 10-year risk of JBS CVD Risk is <0.5 %

The equivalent risk calculation with diastolic BP is <0.5 %

The output for participant one from the old version of Patient UK Website

Cardiovascular Risk Calculator For Primary Prevention

This calculator should not be used if patient has known CVD or diabetes (already known to be at high risk)

Age (30-74)	35	Smoking Status	Smoker
Sex	Male	Glucose	Normal
Systolic BP	135	LVH	No LVH
Diastolic BP	70.5	Central Obesity	No
Total Cholesterol	4.1	South Asian Origin	No
HDL Cholesterol	1.7	Family History of CVD (Men <55 and women <65 years)	No FH
Total /HDL Ratio	2.41	<input type="button" value="Calculate"/> <input type="button" value="Clear Fields"/>	
Serum TG mmol/L			

Using Systolic BP prediction, the 10-year risk of JBS CVD Risk is 2 %

The equivalent risk calculation with diastolic BP is 1 %

The outputs for participant two from the old version of Patient UK Website:

Table 8.1B: The Raw data of the first 30 participants used in Chapter 8:

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	HB	BMI	BPH	BPL	HDL	MAP	DIA	TC	SMK	CVD	ECG
13,956,102.00	No	No	60	Women	White	34.00	13.20	122.50	88.50	1.80	100.00	No	5.20	Yes	No	Yes
63,535,102.00	Yes	Yes	30	Women	White	46.00	13.71	120.00	74.00	1.40	89.50	No	4.50	Yes	Yes	No
71,831,101.00	No	No	66	Women	White	89.00	14.32	159.00	70.00	1.90	99.50	No	6.90	No	No	Yes
34,031,101.00	No	No	84	Women	White	48.50	16.17	112.00	63.50	2.20	80.00	No	5.00	Yes	Yes	Yes
72,604,102.00	No	No	59	Women	White	36.50	16.19	109.50	73.00	2.00	85.00	No	6.00	No	No	No
13,008,101.00	Yes	Yes	50	Women	White	43.00	16.65	117.00	74.00	1.70	88.50	No	6.00	Yes	Yes	Yes
39,139,101.00	No	No	34	Women	White	48.00	16.81	102.00	54.00	1.80	70.00	No	6.50	Yes	No	No
47,856,102.00	No	No	51	Women	White	44.00	16.85	100.50	56.50	1.90	71.00	No	5.10	No	No	No
37,710,101.00	No	No	61	Women	White	43.00	17.43	120.00	77.00	1.20	91.50	No	5.50	No	No	No
54,256,101.00	No	No	31	Women	White	40.00	17.72	124.00	84.00	2.00	97.00	No	3.90	Yes	No	No
53,817,101.00	No	No	43	Women	White	39.00	17.72	107.00	68.00	1.70	81.00	No	4.90	Yes	No	No
44,633,102.00	No	No	39	Women	White	39.00	17.87	106.00	67.00	1.40	80.00	No	5.30	No	No	No
34,523,102.00	No	No	45	Women	White	51.50	18.09	120.00	68.50	2.00	85.50	No	4.70	No	No	No
72,323,102.00	No	No	38	Women	White	39.00	18.10	116.00	77.00	1.40	90.00	No	4.20	Yes	Yes	No
42,831,101.00	No	No	35	Men	White	64.50	18.15	135.00	70.50	1.70	92.00	No	4.10	Yes	No	No
72,833,102.00	No	No	40	Men	White	58.50	18.19	117.00	58.50	1.50	78.00	No	4.60	Yes	No	No
15,047,101.00	No	No	86	Women	White	79.00	18.27	138.00	59.00	2.50	85.50	No	8.90	Yes	No	Yes
23,202,102.00	No	No	58	Women	White	48.00	18.29	113.50	65.50	1.10	81.50	No	3.60	Yes	No	No
20,408,101.00	No	No	44	Women	White	38.00	18.41	100.00	62.00	0.90	75.00	No	4.90	No	No	No
76,910,102.00	No	No	48	Women	White	46.00	18.43	110.00	64.00	2.00	79.00	No	8.00	Yes	No	No
13,856,101.00	No	No	51	Women	White	32.50	18.45	92.50	60.00	2.60	70.50	No	6.70	Yes	No	Yes
21,413,101.00	Yes	Yes	86	Men	White	65.50	18.48	149.00	83.50	1.20	105.00	No	5.60	No	Yes	No
25,713,101.00	No	No	38	Women	White	32.50	18.55	103.50	71.00	1.30	81.50	No	4.00	Yes	No	No
29,646,101.00	No	No	32	Women	White	36.00	18.58	113.50	77.50	1.40	89.50	No	4.60	No	No	No
59,015,101.00	No	No	31	Women	Asian or Asian British	38.00	18.59	100.00	62.00	1.50	74.50	No	4.50	Yes	No	No
48,733,102.00	No	No	48	Women	White	50.50	18.62	133.00	82.50	2.30	99.50	No	6.10	Yes	No	No
51,623,102.00	No	No	49	Women	White	71.00	18.63	168.50	97.50	1.90	121.50	No	4.80	No	No	No
71,718,102.00	No	No	57	Women	White	38.50	18.63	105.00	66.50	2.20	79.00	No	5.80	No	No	Yes
34,846,101.00	No	No	32	Women	White	44.50	18.65	118.50	74.00	1.20	89.00	No	7.00	No	No	Yes
20,023,102.00	No	No	77	Women	White	57.00	18.66	136.50	79.50	2.00	98.50	No	7.20	Yes	No	No

Table 8.2B: Predicative Percentage Risks of 10 years for the first 30 participants based on Internet Model – I NHS BlackHeath centre (ref: <http://www.bhgp.co.uk/chdriskresult.asp>).

Pserial no.	Age	Sex	BMI	BPH	BPL	HDL	DIA	TC	SMK	ECG	% PR
13,956,102.00	60	Women	13.20	122.50	88.50	1.80	No	5.20	Yes	Yes	7 7
63,535,102.00	30	Women	13.71	120.00	74.00	1.40	No	4.50	Yes	No	1 1
71,831,101.00	66	Women	14.32	159.00	70.00	1.90	No	6.90	No	Yes	8 8
34,031,101.00	84	Women	16.17	112.00	63.50	2.20	No	5.00	Yes	Yes	3 NA
72,604,102.00	59	Women	16.19	109.50	73.00	2.00	No	6.00	No	No	3 3
13,008,101.00	50	Women	16.65	117.00	74.00	1.70	No	6.00	Yes	Yes	1 3
39,139,101.00	34	Women	16.81	102.00	54.00	1.80	No	6.50	Yes	No	2 1
47,856,102.00	51	Women	16.85	100.50	56.50	1.90	No	5.10	No	No	11 2
37,710,101.00	61	Women	17.43	120.00	77.00	1.20	No	5.50	No	No	0 11
54,256,101.00	31	Women	17.72	124.00	84.00	2.00	No	3.90	Yes	No	1 0
53,817,101.00	43	Women	17.72	107.00	68.00	1.70	No	4.90	Yes	No	1 1
44,633,102.00	39	Women	17.87	106.00	67.00	1.40	No	5.30	No	No	2 1
34,523,102.00	45	Women	18.09	120.00	68.50	2.00	No	4.70	No	No	1 2
72,323,102.00	38	Women	18.10	116.00	77.00	1.40	No	4.20	Yes	No	3 1
42,831,101.00	35	Men	18.15	135.00	70.50	1.70	No	4.10	Yes	No	(5) 3
72,833,102.00	40	Men	18.19	117.00	58.50	1.50	No	4.60	Yes	No	6 5
15,047,101.00	86	Women	18.27	138.00	59.00	2.50	No	8.90	Yes	Yes	3 NA
23,202,102.00	58	Women	18.29	113.50	65.50	1.10	No	3.60	Yes	No	4 8
20,408,101.00	44	Women	18.41	100.00	62.00	0.90	No	4.90	No	No	3 2
76,910,102.00	48	Women	18.43	110.00	64.00	2.00	No	8.00	Yes	No	1 4
13,856,101.00	51	Women	18.45	92.50	60.00	2.60	No	6.70	Yes	Yes	0 3
21,413,101.00	86	Men	18.48	149.00	83.50	1.20	No	5.60	No	No	0 NA
25,713,101.00	38	Women	18.55	103.50	71.00	1.30	No	4.00	Yes	No	4 1
29,646,101.00	32	Women	18.58	113.50	77.50	1.40	No	4.60	No	No	5 0
59,015,101.00	31	Women	18.59	100.00	62.00	1.50	No	4.50	Yes	No	3 0
48,733,102.00	48	Women	18.62	133.00	82.50	2.30	No	6.10	Yes	No	1 4
51,623,102.00	49	Women	18.63	168.50	97.50	1.90	No	4.80	No	No	10 5
71,718,102.00	57	Women	18.63	105.00	66.50	2.20	No	5.80	No	Yes	7 3
34,846,101.00	32	Women	18.65	118.50	74.00	1.20	No	7.00	No	Yes	2 1
20,023,102.00	77	Women	18.66	136.50	79.50	2.00	No	7.20	Yes	No	1NA

Table 8.3B: Calculation of TPR, FPR, LRP, and LRN, for the Internet Model I for the first 30 Participants NHS BlackHeath centre (ref: <http://www.bhgp.co.uk/chdriskresult.asp>)

Pserial no.	Grp	Bp1	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRM
13,956,102.00	No	No	60	Women	7	0	1	1	0.99968	3125	0
63,535,102.00	Yes	Yes	30	Women	1	0	1	1	0.99936	1562.5	0
71,831,101.00	No	No	66	Women	8	0	1	1	0.999041	1042.753	0
34,031,101.00	No	No	84	Women	3	0	1	1	0.998721	781.8608	0
72,604,102.00	No	No	59	Women	3	0	1	1	0.998401	625.3909	0
13,008,101.00	Yes	Yes	50	Women	1	0	1	1	0.998081	521.1047	0
39,139,101.00	No	No	34	Women	2	0	1	1	0.997761	446.628	0
47,856,102.00	No	No	51	Women	11	0	1	1	0.997442	390.9304	0
37,710,101.00	No	No	61	Women	0	0	1	1	0.997122	347.4635	0
54,256,101.00	No	No	31	Women	1	0	1	1	0.996802	312.6954	0
53,817,101.00	No	No	43	Women	1	0	1	1	0.996482	284.2524	0
44,633,102.00	No	No	39	Women	2	0	1	1	0.996162	260.5524	0
34,523,102.00	No	No	45	Women	1	0	1	1	0.995843	240.5581	0
72,323,102.00	No	No	38	Women	3	0	1	1	0.995523	223.3639	0
42,831,101.00	No	No	35	Men	5	0	1	1	0.995203	208.4636	0
72,833,102.00	No	No	40	Men	6	0	1	1	0.994883	195.427	0
15,047,101.00	No	No	86	Women	3	0	1	1	0.994563	183.925	0
23,202,102.00	No	No	58	Women	4	0	1	1	0.994244	173.7318	0
20,408,101.00	No	No	44	Women	3	0	1	1	0.993924	164.582	0
76,910,102.00	No	No	48	Women	1	0	1	1	0.993604	156.3477	0
13,856,101.00	No	No	51	Women	0	0	1	1	0.993284	148.8982	0
21,413,101.00	Yes	Yes	86	Men	0	0	1	1	0.992965	142.1464	0
25,713,101.00	No	No	38	Women	4	0	1	1	0.992645	135.9619	0
29,646,101.00	No	No	32	Women	5	0	1	1	0.992325	130.2932	0
59,015,101.00	No	No	31	Women	3	0	1	1	0.992005	125.0782	0
48,733,102.00	No	No	48	Women	1	0	1	1	0.991685	120.2646	0
51,623,102.00	No	No	49	Women	10	0	1	1	0.991366	115.8212	0
71,718,102.00	No	No	57	Women	7	0	1	1	0.991046	111.6819	0
34,846,101.00	No	No	32	Women	2	0	1	1	0.990726	107.8283	0
20,023,102.00	No	No	77	Women	1	0	1	1	0.990406	104.2318	0

Table 8.4B: Predicative Percentage Risks for 10 years for the first 30 participants based on Internet Model – II Patient UK User Survey (ref: <http://www.patient.co.uk/doctor/Primary-Cardiovascular-Risk-Calculator.htm>)

Pserial no.	Age	Sex	BMI	BPH	BPL	HDL	DIA	TC	SMK	ECG	% PR
13,956,102.00	60	Women	13.20	122.50	88.50	1.80	No	5.20	Yes	Yes	9 21
63,535,102.00	30	Women	13.71	120.00	74.00	1.40	No	4.50	Yes	No	4.3 0.5
71,831,101.00	66	Women	14.32	159.00	70.00	1.90	No	6.90	No	Yes	20.9 30
34,031,101.00	84	Women	16.17	112.00	63.50	2.20	No	5.00	Yes	Yes	17.5 NA
72,604,102.00	59	Women	16.19	109.50	73.00	2.00	No	6.00	No	No	11.8 4
13,008,101.00	50	Women	16.65	117.00	74.00	1.70	No	6.00	Yes	Yes	9.3 17*
39,139,101.00	34	Women	16.81	102.00	54.00	1.80	No	6.50	Yes	No	8.7 0.5
47,856,102.00	51	Women	16.85	100.50	56.50	1.90	No	5.10	No	No	10 2
37,710,101.00	61	Women	17.43	120.00	77.00	1.20	No	5.50	No	No	7.5 8
54,256,101.00	31	Women	17.72	124.00	84.00	2.00	No	3.90	Yes	No	4.8 0.5
53,817,101.00	43	Women	17.72	107.00	68.00	1.70	No	4.90	Yes	No	6.8 2*
44,633,102.00	39	Women	17.87	106.00	67.00	1.40	No	5.30	No	No	5.8 1
34,523,102.00	45	Women	18.09	120.00	68.50	2.00	No	4.70	No	No	8.6 1
72,323,102.00	38	Women	18.10	116.00	77.00	1.40	No	4.20	Yes	No	4.6 4
42,831,101.00	35	Men	18.15	135.00	70.50	1.70	No	4.10	Yes	No	1.7 2*
72,833,102.00	40	Men	18.19	117.00	58.50	1.50	No	4.60	Yes	No	2.5 2*
15,047,101.00	86	Women	18.27	138.00	59.00	2.50	No	8.90	Yes	Yes	46 NA
23,202,102.00	58	Women	18.29	113.50	65.50	1.10	No	3.60	Yes	No	5.2 8
20,408,101.00	44	Women	18.41	100.00	62.00	0.90	No	4.90	No	No	4.3 3
76,910,102.00	48	Women	18.43	110.00	64.00	2.00	No	8.00	Yes	No	14.4 6
13,856,101.00	51	Women	18.45	92.50	60.00	2.60	No	6.70	Yes	Yes	14.9 11
21,413,101.00	86	Men	18.48	149.00	83.50	1.20	No	5.60	No	No	7.3 NA
25,713,101.00	38	Women	18.55	103.50	71.00	1.30	No	4.00	Yes	No	4.1 1
29,646,101.00	32	Women	18.58	113.50	77.50	1.40	No	4.60	No	No	4.2 0.5
59,015,101.00	31	Women	18.59	100.00	62.00	1.50	No	4.50	Yes	No	4.6 0.5
48,733,102.00	48	Women	18.62	133.00	82.50	2.30	No	6.10	Yes	No	12.1 5
51,623,102.00	49	Women	18.63	168.50	97.50	1.90	No	4.80	No	No	8.9 5
71,718,102.00	57	Women	18.63	105.00	66.50	2.20	No	5.80	No	Yes	12.7 8
34,846,101.00	32	Women	18.65	118.50	74.00	1.20	No	7.00	No	Yes	5.5 3
20,023,102.00	77	Women	18.66	136.50	79.50	2.00	No	7.20	Yes	No	20.2 NA

Table 8.5B: Calculation of TPR, FPR, LRP, and LRN, for the Internet Model – II for the first 30 participants Patient UK User Survey (ref: <http://www.patient.co.uk/doctor/Primary-Cardiovascular-Risk-Calculator.htm>)

Pserial no.	Grp	BpI	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRN
13,956,102.00	No	No	60	Women	9	0	1	1	0.9997	3584.22	0
63,535,102.00	Yes	Yes	30	Women	4.3	0	1	1	0.9994	1788.90	0
71,831,101.00	No	No	66	Women	20.9	1	0	0.984848	0.9994	1761.80	0.01516
34,031,101.00	No	No	84	Women	17.5	0	1	0.984848	0.9991	1175.23	0.01516
72,604,102.00	No	No	59	Women	11.8	0	1	0.984848	0.9988	880.90	0.01516
13,008,101.00	Yes	Yes	50	Women	9.3	0	1	0.984848	0.9986	704.97	0.01517
39,139,101.00	No	No	34	Women	8.7	0	1	0.984848	0.9983	587.61	0.01517
47,856,102.00	No	No	51	Women	10	0	1	0.984848	0.9980	503.50	0.01518
37,710,101.00	No	No	61	Women	7.5	0	1	0.984848	0.9977	440.64	0.01518
54,256,101.00	No	No	31	Women	4.8	0	1	0.984848	0.9974	391.58	0.01519
53,817,101.00	No	No	43	Women	6.8	0	1	0.984848	0.9972	352.48	0.01519
44,633,102.00	No	No	39	Women	5.8	0	1	0.984848	0.9969	320.48	0.01519
34,523,102.00	No	No	45	Women	8.6	0	1	0.984848	0.9966	293.72	0.01520
72,323,102.00	No	No	38	Women	4.6	0	1	0.984848	0.9963	271.15	0.01520
42,831,101.00	No	No	35	Men	1.7	0	1	0.984848	0.9960	251.75	0.01521
72,833,102.00	No	No	40	Men	2.5	0	1	0.984848	0.9958	234.992	0.01521
15,047,101.00	No	No	86	Women	46	1	0	0.969697	0.9958	231.37	0.03043
23,202,102.00	No	No	58	Women	5.2	0	1	0.969697	0.9955	216.889	0.03043
20,408,101.00	No	No	44	Women	4.3	0	1	0.969697	0.9952	204.14	0.03044
76,910,102.00	No	No	48	Women	14.4	0	1	0.969697	0.9949	192.82	0.03045
13,856,101.00	No	No	51	Women	14.9	0	1	0.969697	0.9946	182.655	0.03046
21,413,101.00	Yes	Yes	86	Men	7.3	0	1	0.969697	0.9944	173.53	0.03047
25,713,101.00	No	No	38	Women	4.1	0	1	0.969697	0.9941	165.257	0.03048
29,646,101.00	No	No	32	Women	4.2	0	1	0.969697	0.9938	157.75	0.03049
59,015,101.00	No	No	31	Women	4.6	0	1	0.969697	0.9935	150.90	0.03049
48,733,102.00	No	No	48	Women	12.1	0	1	0.969697	0.9932	144.60	0.03050
51,623,102.00	No	No	49	Women	8.9	0	1	0.969697	0.9930	138.82	0.03051
71,718,102.00	No	No	57	Women	12.7	0	1	0.969697	0.9927	133.47	0.03052
34,846,101.00	No	No	32	Women	5.5	0	1	0.969697	0.9924	128.53	0.03053
20,023,102.00	No	No	77	Women	20.2	1	0	0.954545	0.9924	126.53	0.04580

Table 8.6B: Predicative Percentage Risks of 10 years for the first 30 participants based on Framingham equation model I – II – III (I – USA, II – International, III – UK)

Pserial no.	Grp	Bp1	Age	Sex	Ethnic	USA_PR	INT_PR	UKMEN_P R	UK_PR
13,956,102.00	No	No	60	Women	White	17.58	17.58	0.00	17.58
63,535,102.00	Yes	Yes	30	Women	White	0.13	0.13	0.00	0.13
71,831,101.00	No	No	66	Women	White	22.55	22.55	0.00	22.55
34,031,101.00	No	No	84	Women	White	12.87	12.87	0.00	12.87
72,604,102.00	No	No	59	Women	White	2.82	2.82	0.00	2.82
13,008,101.00	Yes	Yes	50	Women	White	15.12	15.12	0.00	15.12
39,139,101.00	No	No	34	Women	White	0.40	0.40	0.00	0.40
47,856,102.00	No	No	51	Women	White	1.22	1.22	0.00	1.22
37,710,101.00	No	No	61	Women	White	7.02	7.02	0.00	7.02
54,256,101.00	No	No	31	Women	White	0.05	0.05	0.00	0.05
53,817,101.00	No	No	43	Women	White	1.59	1.59	0.00	1.59
44,633,102.00	No	No	39	Women	White	0.66	0.66	0.00	0.66
34,523,102.00	No	No	45	Women	White	0.82	0.82	0.00	0.82
72,323,102.00	No	No	38	Women	White	0.97	0.97	0.00	0.97
42,831,101.00	No	No	35	Men	White	1.74	1.74	1.74	1.74
72,833,102.00	No	No	40	Men	White	3.18	3.18	3.18	3.18
15,047,101.00	No	No	86	Women	White	26.00	26.00	0.00	26.00
23,202,102.00	No	No	58	Women	White	6.27	6.27	0.00	6.27
20,408,101.00	No	No	44	Women	White	2.43	2.43	0.00	2.43
76,910,102.00	No	No	48	Women	White	4.85	4.85	0.00	4.85
13,856,101.00	No	No	51	Women	White	7.37	7.37	0.00	7.37
21,413,101.00	Yes	Yes	86	Men	White	31.65	31.65	31.65	31.65
25,713,101.00	No	No	38	Women	White	0.75	0.75	0.00	0.75
29,646,101.00	No	No	32	Women	White	0.08	0.08	0.00	0.08
59,015,101.00	No	No	31	Women	Asian or Asian British	0.08	0.08	0.00	0.08
48,733,102.00	No	No	48	Women	White	3.69	3.69	0.00	3.69
51,623,102.00	No	No	49	Women	White	3.18	3.18	0.00	3.18
71,718,102.00	No	No	57	Women	White	7.17	7.17	0.00	7.17
34,846,101.00	No	No	32	Women	White	2.37	2.37	0.00	2.37
20,023,102.00	No	No	77	Women	White	11.87	11.87	0.00	11.87

Table 8.7B: Calculation of TPR, FPR, LRP, and LRN, for the for the first 30 participants based on Framingham equation model I – II – III (I – USA, II – International, III – UK)

Pserial no.	Grp	Bp1	Age	Sex	%PR	EX	NEX	TPR	FPR	LRP	LRN
13,956,102.00	No	No	60	Women	17.58	0	1	1	1.00	2808.99	0
63,535,102.00	Yes	Yes	30	Women	0.13	0	1	1	1.00	1406.47	0
71,831,101.00	No	No	66	Women	22.55	1	0	0.9988	0.9993	1404.7820	0.0012
34,031,101.00	No	No	84	Women	12.87	0	1	0.9988	0.9989	936.0825	0.0012
72,604,102.00	No	No	59	Women	2.82	0	1	0.9988	0.9986	702.3910	0.0012
13,008,101.00	Yes	Yes	50	Women	15.12	0	1	0.9988	0.9982	561.7548	0.0012
39,139,101.00	No	No	34	Women	0.40	0	1	0.9988	0.9979	468.0412	0.0012
47,856,102.00	No	No	51	Women	1.22	0	1	0.9988	0.9975	401.2857	0.0012
37,710,101.00	No	No	61	Women	7.02	0	1	0.9988	0.9972	351.0721	0.0012
54,256,101.00	No	No	31	Women	0.05	0	1	0.9988	0.9968	312.0275	0.0012
53,817,101.00	No	No	43	Women	1.59	0	1	0.9988	0.9964	280.8774	0.0012
44,633,102.00	No	No	39	Women	0.66	0	1	0.9988	0.9961	255.3170	0.0012
34,523,102.00	No	No	45	Women	0.82	0	1	0.9988	0.9957	234.0755	0.0012
72,323,102.00	No	No	38	Women	0.97	0	1	0.9988	0.9954	216.0502	0.0012
42,831,101.00	No	No	35	Men	1.74	0	1	0.9988	0.9950	200.6025	0.0012
72,833,102.00	No	No	40	Men	3.18	0	1	0.9988	0.9947	187.2516	0.0012
15,047,101.00	No	No	86	Women	26.00	1	0	0.9976	0.9947	187.0264	0.0024
23,202,102.00	No	No	58	Women	6.27	0	1	0.9976	0.9943	175.3250	0.0024
20,408,101.00	No	No	44	Women	2.43	0	1	0.9976	0.9940	165.0015	0.0024
76,910,102.00	No	No	48	Women	4.85	0	1	0.9976	0.9936	155.8505	0.0024
13,856,101.00	No	No	51	Women	7.37	0	1	0.9976	0.9932	147.6393	0.0024
21,413,101.00	Yes	Yes	86	Men	31.65	1	0	0.9964	0.9932	147.4617	0.0036
25,713,101.00	No	No	38	Women	0.75	0	1	0.9964	0.9929	140.1011	0.0036
29,646,101.00	No	No	32	Women	0.08	0	1	0.9964	0.9925	133.4225	0.0036
59,015,101.00	No	No	31	Women	0.08	0	1	0.9964	0.9922	127.3516	0.0036
48,733,102.00	No	No	48	Women	3.69	0	1	0.9964	0.9918	121.8241	0.0036
51,623,102.00	No	No	49	Women	3.18	0	1	0.9964	0.9915	116.7427	0.0036
71,718,102.00	No	No	57	Women	7.17	0	1	0.9964	0.9911	112.0809	0.0036
34,846,101.00	No	No	32	Women	2.37	0	1	0.9964	0.9908	107.7654	0.0036
20,023,102.00	No	No	77	Women	11.87	0	1	0.9964	0.9904	103.7699	0.0036

Appendix 9.0 for Chapter 9

Table 9.1B: No. of CMAUT and Non-CMAUT constraints for substitutable and complementary

No of organs (x) in the (OR)/(AND) combinatorial	No of constraints (inequalities) (y) for AND with CMAUT	No of constraints (inequalities) (y) for OR with and without CMAUT	No of constraints (inequalities) (y) for AND with non- CMAUT
2	3	4	4
3	4	8	6
4	5	16	8
5	6	32	10
6	7	64	12
7	8	128	14
8	9	256	16
9	10	512	18
10	11	1024	20
11	12	2048	22
12	13	4096	24
13	14	8192	26
14	15	16384	28
15	16	32768	30
16	17	65536	32

Table 9.2B: Data sizes for first 30 participants before and after optimisation with CMAUT model:

Table 9.2C: Data sizes for 402 participants before and after optimisation with CMAUT framework:

Table 9.2B: Data size for first 30 participants before and after optimisation with CMAUT:

No. Of participants	Pserial no.	Data size before optimisation (bytes)	Data size after optimisation (bytes)
1	10102102.00	1191	462
2	10104102.00	1195	462
3	10106101.00	1240	464
4	10117101.00	1213	464
5	10135101.00	1203	463
6	10135102.00	1235	464
7	10147102.00	1227	464
8	10149101.00	1225	463
9	10156102.00	1195	463
10	10206101.00	1220	465
11	10208101.00	1199	463
12	10306101.00	1170	464
13	10314101.00	1208	463
14	10318101.00	1206	465
15	10323101.00	1242	462
16	10333101.00	1215	464
17	10333102.00	1206	463
18	10335101.00	1262	463
19	10408101.00	1243	464
20	10408102.00	1242	464
21	10418101.00	1199	464
22	10418102.00	1223	464
23	10446101.00	1209	463
24	10506102.00	1265	465
25	10608101.00	1218	464
26	10710101.00	1221	466
27	10713101.00	1227	464
28	10806101.00	1219	463
29	10808101.00	1221	465
30	10815102.00	1213	463

Table 9.4B: Comparison of Absolute Percentage Risk values from CMAUT models 1 and 2
using the first 30 participants:

Pserial no.	Grp	Bp1	Age	Sex	%PR(M-I Absol)	%PR(M-II Abso)	%PR(M-I Pre)	%PR(M-II Pre)	%PR(I-I Pr)	%PR(I-II Pre)	%PR(F-I P)
13,956,102.00	No	No	60	Women	14.6	13.8	15.6	14.79	7	9	17.58
63,535,102.00	Yes	Yes	30	Women	15.4	14.6	16.3	15.53	1	4.3	0.13
71,831,101.00	No	No	66	Women	16.7	15.4	17.7	16.42	8	20.9	22.55
34,031,101.00	No	No	84	Women	18.0	16.6	19.0	17.64	3	17.5	12.87
72,604,102.00	No	No	59	Women	18.9	17.3	19.8	18.26	3	11.8	2.82
31,510,102.00	No	No	20	Women	20.1	15.2	21.0	16.13	1	9.3	15.12
18,633,105.00	No	No	16	Women	18.5	19.5	19.4	20.28	2	8.7	0.40
13,008,101.00	Yes	Yes	50	Women	16.3	19.9	17.2	20.75	11	10	1.22
60,417,102.00	No	No	16	Men	17.4	14.3	18.3	15.26	0	7.5	7.02
39,139,101.00	No	No	34	Women	21.4	13.3	22.2	14.26	1	4.8	0.05
47,856,102.00	No	No	51	Women	21.9	17.9	22.8	18.81	1	6.8	1.59
49,147,103.00	No	No	16	Women	19.4	18.1	20.2	19.02	2	5.8	0.66
66,035,102.00	No	No	18	Men	18.3	14.7	19.2	15.65	1	8.6	0.82
80,356,105.00	No	No	17	Men	19.5	15.4	20.4	16.33	3	4.6	0.97
37,710,101.00	No	No	61	Women	15.2	12.9	16.2	13.86	5	1.7	1.74
20,556,101.00	No	No	18	Men	20.7	16.8	21.5	17.68	6	2.5	3.18
64,735,105.00	No	No	16	Women	19.1	14.7	20.0	15.67	3	46	26.00
46,449,104.00	No	No	26	Women	17.3	16.2	18.2	17.18	4	5.2	6.27
66,114,101.00	No	No	24	Women	17.1	19.7	18.0	20.58	3	4.3	2.43
50,956,101.00	No	No	23	Women	17.3	17.2	18.1	18.08	1	14.4	4.85
80,002,104.00	No	No	16	Men	17.5	21.9	18.4	22.76	0	14.9	7.37
54,256,101.00	No	No	31	Women	14.1	12.6	15.0	13.61	0	7.3	31.65
53,817,101.00	No	No	43	Women	19.6	18.7	20.5	19.57	4	4.1	0.75
13,433,101.00	No	No	24	Men	14.5	16.0	15.5	16.91	5	4.2	0.08
42,431,101.00	No	No	27	Women	21.2	19.9	22.0	20.68	3	4.6	0.08
44,633,102.00	No	No	39	Women	19.9	10.7	20.8	11.73	1	12.1	3.69
14,602,104.00	No	No	18	Women	23.0	17.2	23.8	18.14	10	8.9	3.18
77,814,101.00	No	No	28	Women	21.1	18.5	21.9	19.39	7	12.7	7.17
51,506,103.00	No	No	21	Men	15.8	14.6	16.7	15.54	2	5.5	2.37
73,304,103.00	No	No	22	Men	10.8	11.0	11.8	12.01	1	20.2	11.87

Chapter 9.8 Tables:

Table 9.8A: Comparison of 10-years PPR values from CMAUT models, Internet calculators and Framingham equations using the first 10 participants in the Chapter 9:

Table 9.8B: Comparison of 10-years PPR values from CMAUT models, Internet calculators and Framingham equations using the first 30 participants in the Thesis:

Table 9.8C: Comparison of 10-years PPR values from CMAUT models, Internet calculators and Framingham equations for the entire 3645 participants in electronic format:

Table 9.8B: Comparison of 10-years PPR values from CMAUT models, Internet calculators and Framingham equations using the first 30 participants:

Pserial no.	Grp	Bp1	Age	Sex	%PR(M-I Absol)	%PR(M-II Abso)	%PR(M-I Pre)	%PR(M-II Pre)	%PR(I-I Pr)	%PR(I-II Pre)	%PR(F-I P)
13,956,102.00	No	No	60	Women	14.6	13.8	15.6	14.79	7	9	17.58
63,535,102.00	Yes	Yes	30	Women	15.4	14.6	16.3	15.53	1	4.3	0.13
71,831,101.00	No	No	66	Women	16.7	15.4	17.7	16.42	8	20.9	22.55
34,031,101.00	No	No	84	Women	18.0	16.6	19.0	17.64	3	17.5	12.87
72,604,102.00	No	No	59	Women	18.9	17.3	19.8	18.26	3	11.8	2.82
31,510,102.00	No	No	20	Women	20.1	15.2	21.0	16.13	1	9.3	15.12
18,633,105.00	No	No	16	Women	18.5	19.5	19.4	20.28	2	8.7	0.40
13,008,101.00	Yes	Yes	50	Women	16.3	19.9	17.2	20.75	11	10	1.22
60,417,102.00	No	No	16	Men	17.4	14.3	18.3	15.26	0	7.5	7.02
39,139,101.00	No	No	34	Women	21.4	13.3	22.2	14.26	1	4.8	0.05
47,856,102.00	No	No	51	Women	21.9	17.9	22.8	18.81	1	6.8	1.59
49,147,103.00	No	No	16	Women	19.4	18.1	20.2	19.02	2	5.8	0.66
66,035,102.00	No	No	18	Men	18.3	14.7	19.2	15.65	1	8.6	0.82
80,356,105.00	No	No	17	Men	19.5	15.4	20.4	16.33	3	4.6	0.97
37,710,101.00	No	No	61	Women	15.2	12.9	16.2	13.86	5	1.7	1.74
20,556,101.00	No	No	18	Men	20.7	16.8	21.5	17.68	6	2.5	3.18
64,735,105.00	No	No	16	Women	19.1	14.7	20.0	15.67	3	46	26.00
46,449,104.00	No	No	26	Women	17.3	16.2	18.2	17.18	4	5.2	6.27
66,114,101.00	No	No	24	Women	17.1	19.7	18.0	20.58	3	4.3	2.43
50,956,101.00	No	No	23	Women	17.3	17.2	18.1	18.08	1	14.4	4.85
80,002,104.00	No	No	16	Men	17.5	21.9	18.4	22.76	0	14.9	7.37
54,256,101.00	No	No	31	Women	14.1	12.6	15.0	13.61	0	7.3	31.65
53,817,101.00	No	No	43	Women	19.6	18.7	20.5	19.57	4	4.1	0.75
13,433,101.00	No	No	24	Men	14.5	16.0	15.5	16.91	5	4.2	0.08
42,431,101.00	No	No	27	Women	21.2	19.9	22.0	20.68	3	4.6	0.08
44,633,102.00	No	No	39	Women	19.9	10.7	20.8	11.73	1	12.1	3.69
14,602,104.00	No	No	18	Women	23.0	17.2	23.8	18.14	10	8.9	3.18
77,814,101.00	No	No	28	Women	21.1	18.5	21.9	19.39	7	12.7	7.17
51,506,103.00	No	No	21	Men	15.8	14.6	16.7	15.54	2	5.5	2.37
73,304,103.00	No	No	22	Men	10.8	11.0	11.8	12.01	1	20.2	11.87

1. Table 9.9A: Comparison of TPR and FPR for CMAUT models, Internet calculators and Framingham equations of the first 10 participants:
2. Table 9.9B: Comparison of TPR and FPR for CMAUT models, Internet calculators and Framingham equations of the first 30 participants in the Thesis:
3. Table 9.9C: Comparison of TPR and FPR for CMAUT models, Internet calculators and Framingham equations for the entire 3645 participants in electronic format;

Table 9.9B: Comparison of TPR and FPR for CMAUT models, Internet calculators and Framingham equations of the first 30 participants:

Pserial no.	Grp	TPR(M-I)	FPR(M-I)	TPR(M-II)	FPR(M-II)	TPR(I-I)	FPR(I-I)	TPR(I-II)	FPR(I-II)	TPR(F-I)	FPR(F-I)
13,956,102.00	No	1	1.000	1	0.9997	1	0.99968	1	0.9997	1	1.00
63,535,102.00	Yes	1	0.999	1	0.9994	1	0.99936	1	0.9994	1	1.00
71,831,101.00	No	1	0.999	1	0.9991	1	0.99904	0.984848	0.9994	0.9988	0.9993
34,031,101.00	No	1	0.999	1	0.9989	1	0.99872	0.984848	0.9991	0.9988	0.9989
72,604,102.00	No	1	0.999	1	0.9986	1	0.99840	0.984848	0.9988	0.9988	0.9986
31,510,102.00	No	0.9983	0.9987	1	0.9983	1	0.99808	0.984848	0.9986	0.9988	0.9982
18,633,105.00	No	0.9983	0.9984	0.9935	0.9983	1	0.99776	0.984848	0.9983	0.9988	0.9979
13,008,101.00	Yes	0.9983	0.9981	0.9871	0.9983	1	0.99744	0.984848	0.9980	0.9988	0.9975
60,417,102.00	No	0.9983	0.9979	0.9871	0.9980	1	0.99712	0.984848	0.9977	0.9988	0.9972
39,139,101.00	No	0.9966	0.9979	0.9871	0.9977	1	0.99680	0.984848	0.9974	0.9988	0.9968
47,856,102.00	No	0.9950	0.9979	0.9871	0.9974	1	0.99648	0.984848	0.9972	0.9988	0.9964
49,147,103.00	No	0.9933	0.9979	0.9871	0.9971	1	0.99616	0.984848	0.9969	0.9988	0.9961
66,035,102.00	No	0.9933	0.9976	0.9871	0.9968	1	0.99584	0.984848	0.9966	0.9988	0.9957
80,356,105.00	No	0.9916	0.9976	0.9871	0.9966	1	0.99552	0.984848	0.9963	0.9988	0.9954
37,710,101.00	No	0.9916	0.9973	0.9871	0.9963	1	0.99520	0.984848	0.9960	0.9988	0.9950
20,556,101.00	No	0.9899	0.9973	0.9871	0.9960	1	0.99488	0.984848	0.9958	0.9988	0.9947
64,735,105.00	No	0.9899	0.9970	0.9871	0.9957	1	0.99456	0.969697	0.9958	0.9976	0.9947
46,449,104.00	No	0.9899	0.9968	0.9871	0.9954	1	0.99424	0.969697	0.9955	0.9976	0.9943
66,114,101.00	No	0.9899	0.9965	0.9806	0.9954	1	0.99392	0.969697	0.9952	0.9976	0.9940
50,956,101.00	No	0.9899	0.9962	0.9806	0.9951	1	0.99360	0.969697	0.9949	0.9976	0.9936
80,002,104.00	No	0.9899	0.9960	0.9742	0.9951	1	0.99328	0.969697	0.9946	0.9976	0.9932
54,256,101.00	No	0.9899	0.9957	0.9742	0.9948	1	0.99296	0.969697	0.9944	0.9964	0.9932
53,817,101.00	No	0.9882	0.9957	0.9742	0.9946	1	0.99264	0.969697	0.9941	0.9964	0.9929
13,433,101.00	No	0.9882	0.9954	0.9742	0.9943	1	0.99232	0.969697	0.9938	0.9964	0.9925
42,431,101.00	No	0.9866	0.9954	0.9677	0.9943	1	0.99200	0.969697	0.9935	0.9964	0.9922
44,633,102.00	No	0.9849	0.9954	0.9677	0.9940	1	0.99168	0.969697	0.9932	0.9964	0.9918
14,602,104.00	No	0.9832	0.9954	0.9677	0.9937	1	0.99136	0.969697	0.9930	0.9964	0.9915
77,814,101.00	No	0.9815	0.9954	0.9677	0.9934	1	0.99104	0.969697	0.9927	0.9964	0.9911
51,506,103.00	No	0.9815	0.9952	0.9677	0.9931	1	0.99072	0.969697	0.9924	0.9964	0.9908
73,304,103.00	No	0.9815	0.9949	0.9677	0.9928	1	0.99040	0.954545	0.9924	0.9964	0.9904

1. Table 9.10A: Comparison of LRP and LRN for CMAUT models, Internet calculators and Framingham equations for the first 10 participants in the Thesis;
2. Table 9.10B: Comparison of LRP and LRN for CMAUT models, Internet calculators and Framingham equations for the first 30 participants below;
3. Table 9.10C: Comparison of LRP and LRN for CMAUT models, Internet calculators and Framingham equations for the entire 3645 participants in electronic format;

Table 9.10B: Comparison of LRP and LRN for CMAUT models, Internet calculators and Framingham equations for the first 30 participants;

Pserial no.	Grp	LRPM(I)	LRN(M-I)	LRP(M-II)	LRN(M-II)	LRPI-I	LRN(I-I)	LRP(I-II)	LRN(I-II)	LRP(F-I)	LRN(F-I)
13,956,102.00	No	3484.32	0	3717.472	0	3125	0	3584.22	0	2808.99	0
63,535,102.00	Yes	1745.20	0	1862.197	0	1562.5	0	1788.90	0	1406.47	0
71,831,101.00	No	1162.79	0	1240.695	0	1042.753	0	1761.80	0.01516	1404.7820	0.0012
34,031,101.00	No	872.60	0	930.233	0	781.8608	0	1175.23	0.01516	936.0825	0.0012
72,604,102.00	No	697.84	0	744.048	0	625.3909	0	880.90	0.01516	702.3910	0.0012
31,510,102.00	No	581.73	0	742.7969	0.0017	521.1047	0	704.97	0.01517	561.7548	0.0012
18,633,105.00	No	577.98	0.0065	619.3046	0.0017	446.628	0	587.61	0.01517	468.0412	0.0012
13,008,101.00	Yes	574.23	0.0129	530.7384	0.0017	390.9304	0	503.50	0.01518	401.2857	0.0012
60,417,102.00	No	492.07	0.0129	464.3344	0.0017	347.4635	0	440.64	0.01518	351.0721	0.0012
39,139,101.00	No	430.67	0.0129	463.5530	0.0034	312.6954	0	391.58	0.01519	312.0275	0.0012
47,856,102.00	No	382.74	0.0129	462.7712	0.0051	284.2524	0	352.48	0.01519	280.8774	0.0012
49,147,103.00	No	344.54	0.0129	461.9893	0.0067	260.5524	0	320.48	0.01519	255.3170	0.0012
66,035,102.00	No	313.17	0.0129	410.6147	0.0067	240.5581	0	293.72	0.01520	234.0755	0.0012
80,356,105.00	No	287.11	0.0129	409.9202	0.0084	223.3639	0	271.15	0.01520	216.0502	0.0012
37,710,101.00	No	264.99	0.0130	369.0350	0.0084	208.4636	0	251.75	0.01521	200.6025	0.0012
20,556,101.00	No	246.10	0.0130	368.4094	0.0101	195.427	0	234.992	0.01521	187.2516	0.0012
64,735,105.00	No	229.66	0.0130	334.8836	0.0101	183.925	0	231.37	0.03043	187.0264	0.0024
46,449,104.00	No	215.29	0.0130	306.9507	0.0101	173.7318	0	216.889	0.03043	175.3250	0.0024
66,114,101.00	No	213.88	0.0194	283.3188	0.0101	164.582	0	204.14	0.03044	165.0015	0.0024
50,956,101.00	No	201.32	0.0194	263.1356	0.0101	156.3477	0	192.82	0.03045	155.8505	0.0024
80,002,104.00	No	200.00	0.0259	245.5758	0.0101	148.8982	0	182.655	0.03046	147.6393	0.0024
54,256,101.00	No	188.87	0.0259	230.2130	0.0101	142.1464	0	173.53	0.03047	147.4617	0.0036
53,817,101.00	No	178.95	0.0259	229.8221	0.0118	135.9619	0	165.257	0.03048	140.1011	0.0036
13,433,101.00	No	169.99	0.0260	216.2913	0.0118	130.2932	0	157.75	0.03049	133.4225	0.0036
42,431,101.00	No	168.86	0.0324	215.9236	0.0135	125.0782	0	150.90	0.03049	127.3516	0.0036
44,633,102.00	No	160.83	0.0325	215.5557	0.0152	120.2646	0	144.60	0.03050	121.8241	0.0036
14,602,104.00	No	153.51	0.0325	215.1878	0.0169	115.8212	0	138.82	0.03051	116.7427	0.0036
77,814,101.00	No	146.85	0.0325	214.8201	0.0186	111.6819	0	133.47	0.03052	112.0809	0.0036
51,506,103.00	No	140.72	0.0325	202.9177	0.0186	107.8283	0	128.53	0.03053	107.7654	0.0036
73,304,103.00	No	135.10	0.0325	192.2274	0.0186	104.2318	0	126.53	0.04580	103.7699	0.0036

Figure 9.10.x:

This was computed in MATLAB software for each model and the outcomes are shows in the figures below based on the PPR results from the calculation conducted in chapter 5 and 7.

$$y = 4.5 * x^{\{5\}} - 10 * x^{\{4\}} + 7.2 * x^{\{3\}} - 2.7 * x^{\{2\}} + 2 * x - 0.0062$$

Percentage = 64.2712500000000 22.2047182838361

Convert into ratio *format* = 0.22; this is fail

Figure 9.10.1. CMAUT diagnosis model I chapter 5

$$y = -1.2e + 002 * x^{\{7\}} + 4.3e + 002 * x^{\{6\}} - 5.7e + 002 * x^{\{5\}} + 3.7e + 002 * x^{\{4\}} - 1.1e + 002 * x^{\{3\}} + 13 * x^{\{2\}} + 0.98 * x + 0.042$$

Percentage = 284.8755000000000 82.4484731049178 (per)

Convert into ratio format = 0.82 this is excellent

Figure 9.10. 2. CMAUT diagnosis model II chapter 5

From Figure 9.10.1 and 9.10.2 calculations, it is subsumed that CMAUT diagnosis model II has an excellent Prediction accuracy of 0.82 while model I is a poor predictor.

$$y(i) = -33 * x^{\{6\}} + 1e + 002 * x^{\{5\}} - 1.2e + 002 * x^{\{4\}} + 64 * x^{\{3\}} - 16 * x^{\{2\}} + 3.3 * x + 0.0056$$

Percentage = 25.9763500000000 92.4827776034743 (per)

Convert into ratio format = 0. 92; this is excellent

Figure 9.10.3. CMAUT 10 years Prognosis model I chapter 7

$$y = 1.6 * x^{\{8\}} + 24 * x^{\{7\}} - 1.2e + 002 * x^{\{6\}} + 2.3e + 002 * x^{\{5\}} - 2e + 002 * x^{\{4\}} + 93 * x^{\{3\}} - 21 * x^{\{2\}} + 3.6 * x + 0.0008$$

Percentage = 112.0080000000000 55.3603314048996 (per)

Convert into ratio format = 0. 55 this is fair because it is approx. 0.60

Figure 9.10.4. CMAUT 10 years Prognosis model II chapter 7

From Figures 9.10.3 and 9.10.4, it is inferred that CMAUT Prognosis model I has an excellent prediction accuracy of 0.92 but model II is just satisfactory predictor.

$$y(i) = 16 * x^5 - 40 * x^4 + 35 * x^3 - 13 * x^2 + 2.7 * x + 0.027;$$

Percentage = -45.9500000000000 -8.81392818280749 (per)

Convert into ratio format = - 0.088 this is fail because it is negative and less than 0.50

Figure 9.10.5. Internet CVD model I chapter 8

$$y(i) = 9.3 * x^5 - 22 * x^4 + 19 * x^3 - 7.3 * x^2 + 1.7 * x - 0.019$$

Percentage = 29.8412500000000 67.5533028944832 (per)

Convert into ratio format = 0.675 this is fair because it is positive and approx. 0.70

Figure 9.10.6. Internet CVD model II chapter 8

From Figures 9.10.5 and 9.10.6 it is inferred that Internet CVD model I has failed as a predictor because it has a prediction accuracy of 0.08 less than 0.5 and according to Figure 9.9 above it lays negative quadrant but Internet CVD model II is fair predictor.

$$y = -99 * x^{\{7\}} + 3.7e + 002 * x^{\{6\}} - 5.2e + 002 * x^{\{5\}} + 3.5e + 002 * x^{\{4\}} - 1.1e + 002 * x^{\{3\}} + 14 * x^{\{2\}} + 0.91 * x + 0.0028$$

Percentage = -58.0697250000001 -13.8966130802240 (per)

Convert into ratio format = -0.1389 this is fail because it is negative and less than 0.50.

Figure 9.10.7: Framingham equation UK_USA